Use of audio and bioimpedance signals with application of machine learning in classification of laryngeal pathologies

UNIVERSITY OF
WEST LONDON

Doctoral Thesis

Use of Audio and Bioimpedance Signals with Application of Machine Learning in Classification of Laryngeal Pathologies

Julia Zofia Tomaszewska, MSc, BA (Hons)

A thesis submitted in partial fulfilment of the requirements
of the University of West London for the degree of Doctor of Philosophy.

School of Computing and Engineering
University of West London

May 2025

ABSTRACT

The early and accurate detection of laryngeal pathologies, with particular focus on malignant lesions, remains a major clinical challenge due to the limitations and invasiveness of conventional diagnostic methods. This thesis addresses this challenge by developing a non-invasive, robust, and accurate classification framework, capable of identifying the cancerous and precancerous lesions with high precision and sensitivity, based on the combined input of audio recordings of human phonation and simultaneous laryngeal bioimpedance measurements obtained through electroglottography (EGG). The signals form a custom dataset developed specifically for the purposes of this study.

The unimodal deep learning classification architectures are developed and rigorously evaluated for each data modality. The highest-performing models are then used as modality-specific building blocks for the final multimodal classification system. The developed multimodal classifiers consistently outperform the unimodal baselines, particularly in prioritising the detection of cancerous and precancerous lesions. This accuracy is further enhanced using feature extraction methods based on the Equivalent Rectangular Bandwidth spectrum, which outperform alternative feature representations. Additionally, continuous speech provides a richer and more discriminative feature set than sustained phonation, leading to improved classification performance.

The final multimodal system, applied through late fusion stacked generalisation, combines the input of audio-derived GTCCs and EGG-derived Gammatone spectrograms, processed through separate one-dimensional Convolutional Neural Networks (CNNs), and integrated at the decision level using an ECOC-based meta-classifier. This system achieves the highest performance metrics, with average results over 10-fold cross-validation of 94.92% ± 2.82% accuracy, 96.67% ± 2.90% precision, 93.07% ± 3.53% sensitivity, 96.77%

± 2.84% specificity, and 94.81% ± 2.90% F1 score for pathology detection, and the strongest performance in multi-class classification, particularly for malignant case detection (89.23% ± 1.95 accuracy, 82.32% ± 2.78 precision, 83.27% ± 3.43 sensitivity, 91.90% ± 1.75 specificity, and 82.76% ± 2.6 F1).

This work advances the understanding of multimodal laryngeal pathology classification and lays the foundation for future research into clinically deployable and real-time pathology screening tools.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AC	Alternating Current
ANOVA	Analysis of Variance
Bi-LSTM	Bi-directional Long Short Term Memory Network
CNN	Convolutional Neural Network
CI	Confidence Intervals
DAT	Domain Adversarial Training
DC	Direct Current
DCT	Discrete Cosine Transform
dEGG	Differentiated EGG
DL	Deep Learning
DNN	Deep Neural Network
ECOC	Error-Correcting Output Codes
EDA	Exploratory Data Analysis
EGG	Electroglottography
EL	Ensemble Learning classifier
FIR	Finite Impulse Response
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Models
GOI	Glottal Opening Instant
GTCC	Gammatone Spectral Coefficients
HNR	Harmonic to Noise Ratio
HP	High-Pass
HR	Harmonic Ratio
HSD	Honestly Significant Difference

IIR	Infinite Impulse Response
LDA	Linear Discriminant Analysis
LP	Low-Pass
LSTM	Long Short Term Memory Network
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multilayer Perceptron
PC	Principal Component
PCA	Principal Component Analysis
RF	Random Forest classifier
RNN	Recurrent Neural Network
SD	Standard Deviation
SNR	Signal to Noise Ratio
STFT	Short-Time Fourier Transform
SVD	Saarbruecken Voice Database
SVM	Support Vector Machine
THD	Total Harmonic Distortion
VFCA	Vocal Fold Contact Area

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor – Dr Apostolos Georgakis. From the very beginning of this unusual PhD journey, he approached me and this project with commitment like no other. I am especially grateful for many (many!) hours he devoted to our meetings, for all his support, and most importantly, his belief. His knowledge and experience are truly remarkable. I consider myself incredibly fortunate to have had him as my supervisor – his guidance shaped not only my research, but also my development as a researcher.

I would also like to thank my second supervisor, Dr Christos Chousidis, who is somewhat to blame for my PhD in the first place! – and I say that with genuine fondness and gratitude. His encouragement and optimism were vital, and I will always be thankful for the opportunities and perspective he brought to my work.

I am deeply grateful to Professor Massoud Zolgharni, whose generosity in providing access to GPU resources made this research possible. I truly appreciate his support, and belief in the value of this project throughout.

I would like to express my utmost gratitude and sincere thanks to Dr Ewa Migacz and, most importantly, Dr Wojciech Kukwa, for their contributions from Otorhinolaryngology Clinic of the Medical University of Warsaw in the Czerniakowski Hospital, Warsaw, Poland. Without Dr Wojtek, this research simply would not have happened – his support made the data collection possible, and his guidance provided the essential medical perspective that gave this work real-world grounding. He also opened my eyes to what research truly is, and I remain deeply grateful for the opportunities he has created for me. His passion for science is contagious, and I genuinely admire the depth of his knowledge, his enthusiasm for discovery, and the way he manages to keep a sense of humour along the way.

To Nino Auricchio and Dr Gerard Roma – thank you for creating such a supportive and enjoyable academic environment. Your knowledge and experience were invaluable, and your presence made the university feel like home.

To my fellow PhD companions – particularly Daniel Tweddle, and most especially Nicola Russo – thank you for sharing this experience. Nico, your unspoken solidarity, and ongoing support during the final months made a real difference.

But most special thanks go to Dr Eugenio Donati – academic, colleague, and true friend. Without your support, advice, and steady encouragement, I genuinely don't think I would have made it through.

To my most incredible friends Nina and Klaudia, and to my wonderful Ealing crew – thank you for bringing humour, energy, and balance when it was needed most. I would also like to thank Dr Steve Oxnard – your presence mattered more than words can say.

And finally, to the true anchors of my life – my Mum, my Dad, and my Sister – your support has carried me through. Despite the miles between us, you always manage to be the closest. I am deeply grateful for everything you've done to help me reach this point, making me who I am.

"What makes the desert beautiful is that somewhere it hides a well" (*The Little Prince*). I was beyond lucky to have found quite a few of them along the way.

AUTHOR'S DECLARATION

This research has been conducted as part of a PhD at the University of West London under the Vice Chancellor Scholarship Scheme. The material contained in this thesis has not previously been submitted for the achievement of an academic award in this or any other institution. The work presented is the author's own work unless otherwise referenced.

Chapter 1

Introduction

Laryngeal pathologies encompass a diverse spectrum of diseases affecting the structure and function of the larynx – an organ often referred to as the “voice box”, essential for voice production, respiration, and airway protection during swallowing. These conditions range from mild issues, such as vocal fold nodules or throat inflammation, to severe, life-threatening diseases, including various malignant lesions and cancerous growths. Unlike general “voice disorders” – which may also include speech impediments arising from psychological or neurological causes – the laryngeal pathologies refer specifically to physical abnormalities or impairments within the laryngeal structure. Furthermore, while benign voice disorders often manifest as transient hoarseness or vocal fatigue, laryngeal pathologies – especially any malignant lesions – may lead to persistent dysphonia, airway obstruction, and systemic complications. This distinction is therefore critical as the latter demands rapid medical intervention, often involving invasive diagnostic techniques and precise therapeutic approaches. The importance of recognising and diagnosing these conditions is paramount, particularly for malignant and precancerous diseases, where an early intervention significantly improves patient outcomes.

This doctoral thesis focuses on the investigation and identification of reliable digital / computational methods for timely detection and accurate classification of laryngeal pathologies. With the application of machine learning methods, we propose a robust pathology detection and classification systems based on a multimodal approach integrating two types of digital signals; audio and simultaneously recorded laryngeal bioimpedance (collected using an electroglottograph). By leveraging advanced signal processing and deep learning techniques, this research aims to develop a robust diagnostic framework with

particular emphasis on distinguishing healthy individuals, patients with vocal fold paralysis, and those with cancerous or precancerous growths. The work contributes to the growing field of non-invasive diagnostic tools, addressing the critical need for accurate and accessible solutions for early detection of laryngeal disorders.

This chapter contributes the introduction to this work. In the following sections we first introduce the subject, then we explain the rationale and establish the hypothesis of this research, followed by highlighting its aims and objectives. Subsequently, the contributions to knowledge are listed, supported by the publications arising from this study. Finally, we describe the structure of this thesis with a summary of the contents of each chapter.

1. INTRODUCTION

1.1. RESEARCH MOTIVATION

The prevalence of laryngeal pathologies underscores their significance as a public health concern. Studies estimate that nearly 29.9% of individuals experience voice disorders at some point in their lifetime, with laryngeal conditions comprising a significant proportion (Roy *et al.*, 2005). A more recent estimate in 2019 indicated a reduced prevalence of 16.9% (Lyberg-Ahlander *et al.*, 2019). These figures, however, may underestimate the true prevalence, given the lasting effects of factors such as the COVID-19 pandemic, which has been associated with long-term vocal impairments, including persistent hoarseness, dysphonia, or chronic cough (National Health Service, 2022, 2023). Within this context, malignant and precancerous laryngeal pathologies stand out due to their critical impact on survival rates and quality of life, necessitating timely and accurate diagnostic approaches.

Current diagnostic approaches for laryngeal pathologies rely heavily on invasive procedures such as laryngoscopy (Rosen and Murry, 2000) and stroboscopy (Kitzing, 1985). These methods, which involve the insertion of an endoscope to visualise the vocal folds, often require significant patient cooperation and can cause severe discomfort. Additionally, their accuracy is contingent on the expertise of the clinician performing the examination, making the process inherently subjective and prone to variability. Advanced techniques such as direct laryngoscopy (often also referred to as directoscopy) under general anaesthesia may be employed to obtain tissue samples for histopathological analysis, however, these procedures are costly and resource intensive. Thus, there is an urgent need for non-invasive, objective, and cost-effective alternatives for diagnosing laryngeal pathologies.

Recent research recognised the need for an automated vocal tract pathology diagnostic system resulting in numerous studies attempting to classify such disorders based on audio

signals. In recent years, computational methods leveraging machine learning have emerged as promising tools for laryngeal pathology detection and classification; the experiments focus mostly on audio recordings of sustained vowels (Markaki and Stylianou, 2011; Al-Nasheri *et al.*, 2017; Kuo *et al.*, 2023) – with a few investigating the continuous speech (Arias-Londoño *et al.*, 2010; Wang *et al.*, 2022) – as well as audio-derived features, such as Mel-Frequency Cepstral Coefficients (Wang *et al.*, 2022), pitch and amplitude perturbation features (Moran *et al.*, 2006), harmonic-to-noise ratio (Arias-Londoño *et al.*, 2010), and Modulation Spectral Features (Markaki and Stylianou, 2011). Although some achieve accuracy as high as 90%, most vocal tract pathology audio classifiers are limited to binary classification between healthy and pathological signals, or discrimination between two large collective groups of vocal tract diseases (Kuo *et al.*, 2023). Most research often fails to distinguish more specific conditions like cancerous and precancerous growths.

Alternatively, some studies investigate the classification of laryngeal disorders of which symptoms are audible during usual patient auscultation commonly performed by a medical practitioner (Al-Nasheri *et al.*, 2017). An instance of such case is differentiation between vocal fold nodules, usually characterised by an intensely hoarse voice, and vocal fold paralysis, which manifests itself in partial or complete lack of voice or its abnormal weakness, resulting in a breathy sound. Naturally, most reported systems that rely on the application of audio signals and audio-derived features still struggle with the differentiation between disorders that cause similar changes in the sound of the human voice – for instance, functional dysphonia and vocal fold paralysis. In these cases, an alternative modality to audio recordings may help to improve the laryngeal pathology classification results.

Electroglottography (EGG), a non-invasive technique that measures laryngeal bioimpedance during vocal fold vibration, has also been explored as a diagnostic tool.

Although its clinical use has declined due to signal quality and reliability concerns (Colton and Conture, 1990; Herbst and Dunn, 2019), advancements in signal processing and machine learning may revitalise its application. Furthermore, collection of electroglottographic measurements does not require a controlled environment – unlike audio recording – and can be done with low-cost, portable devices. By capturing physiological data that complements audio features, EGG could enhance the accuracy and robustness of machine learning-based laryngeal pathology classification systems. A systematic review of electroglottography-based medical diagnostics was presented in (Tomaszewska and Georgakis, 2023).

However, machine learning and deep learning models require large, high-quality datasets to achieve robust and reliable performance. Unfortunately, the number of publicly available datasets containing both audio and electroglottographic (EGG) signals for laryngeal pathology classification is very limited. Some databases contain various recordings of the same participants referred to by different unique identification numbers. This repetition introduces speaker-dependent features, which can artificially inflate classification accuracy and reduce the generalisability of the classification models. Furthermore, existing datasets are often limited in quality and size – especially with regards to more specific conditions. This lack of standardised, comprehensive datasets poses a major hurdle to advancing computational methods for laryngeal pathology classification.

To address this critical gap, a part of this research focuses on the development of a new dataset comprising synchronised audio and EGG signals collected from participants with various laryngeal pathologies. By ensuring rigorous quality control and accurate labelling, the dataset aims to serve as a reliable resource for future research.

Lastly, while most of research focuses on Mel spectrum-derived feature extraction approaches (Borsky *et al.*, 2017; Muhammad and Alhussain, 2021; Islam *et al.*, 2022;

Miliaresi *et al.*, 2022; Geng *et al.*, 2022), we hypothesise that feature extraction methods derived from the Equivalent Rectangular Bandwidth (ERB) spectrum may improve the performance of laryngeal pathology classification. The ERB spectrum closely models the frequency selectivity of the human auditory system, which can be particularly relevant for capturing subtle differences in pathological signals. Furthermore, we hypothesise that the use of continuous speech instead of sustained phonation in a machine learning-based laryngeal pathology classification model may improve its performance. There is a considerable lack of reporting on the effectiveness of continuous speech in laryngeal pathology classification; the majority of existing systems rely on audio recordings of sustained vowel phonation, especially vowel /a/ (Godino-Llorente and Gomez-Vilda, 2004; Henríquez *et al.*, 2009; Arias-Londoño *et al.*, 2010; Markaki and Stylianou, 2011; Hemmerling *et al.*, 2016; Al-Nasheri *et al.*, 2017; Harar *et al.*, 2017; Borsky *et al.*, 2017; Zhou *et al.*, 2022). This preference is attributed to the stable positioning of the epiglottis, consistent fundamental frequency, and the absence of complex articulatory movements related to the language (Rosa *et al.*, 1999; Henríquez *et al.*, 2009; Markaki and Stylianou, 2011; Islam *et al.*, 2022). Nevertheless, the fluctuating articulatory movements and variations in pitch and intensity in continuous speech may offer a more comprehensive assessment of the larynx under realistic speaking conditions. We argue that the application of continuous speech in machine learning-based laryngeal pathology classification may reveal pathological markers that sustained phonation fails to expose.

This thesis aims to explore the combined potential of audio and EGG signals for the detection and classification of laryngeal pathologies in a multimodal machine learning approach. The effectiveness and accuracy of laryngeal pathology detection and classification based on audio and EGG as singular modalities is also investigated. We conduct a thorough data analysis of audio and simultaneously recorded laryngeal

bioimpedance signals collected during sustained vowel phonation, as well as continuous speech – the dataset created for the purposes of this research. The data analysis investigates statistical and time-domain features, as well as spectral frequency-domain characteristics. By analysing data collected from healthy individuals and patients with a range of laryngeal conditions, this work seeks to identify the most effective computational methods for feature extraction and classification of such signals. Primarily, we focus on the appropriate classification of cancerous and precancerous lesions, neuromuscular disorders, and healthy cases.

The feature extraction techniques, including Mel-Spectrograms, Mel-frequency cepstral Coefficients (MFCCs), Gammatone Cepstral Coefficients (GTCCs) and Gammatone spectrograms are explored to choose the method providing the highest classification accuracy. We seek to prove that the application of feature extraction methods derived from the ERB spectrum will increase the model's performance. For the classification task, we investigate various machine learning methods, including Random Forest (RF), one-dimensional Convolutional Neural Networks (1D-CNN), two-dimensional Convolutional Neural Networks (2D-CNN), Long-Short Term Memory Networks (LSTM), as well as Bi-directional Long-Short Term Memory Networks (BiLSTMs). The performance of all proposed classification systems is examined on both phonation types – the continuous speech, as well as sustained phonation.

Based on the methodology yielding the highest accuracy, a multimodal laryngeal pathology detection and classification system is developed. For the purposes of finding the most appropriate approach to multimodality, three data fusion strategies are evaluated; early fusion, late fusion, and a hybrid – "intermediate", "middle" or "halfway" fusion (Gadzicki *et al.*, 2020).

All proposed classification approaches are tested on two datasets; the custom dataset created specifically for the purposes of this study, as well as Saarbruecken Voice Database (SVD), which is the commonly employed dataset within the field of laryngeal pathology classification (Al-Nasheri *et al.*, 2017; Harar *et al.*, 2017; Zhou *et al.*, 2022; Lee, 2021). This was done to ensure the generalisability of this study's results among various datasets, as well as for the cross-validation purposes.

The ultimate goal is to design a reliable, accurate, and multimodal diagnostic prototype that integrates audio and bioimpedance measurements, offering a rapid and non-invasive alternative for clinical methods.

1.2. RESEARCH RATIONALE

This research investigates the applicability of machine learning and multimodal deep learning for the classification of laryngeal pathologies using audio and laryngeal bioimpedance (EGG) signals. As discussed in the research motivation, the justification for this study arises from the following key considerations:

- Current limitations of traditional diagnostic methods. Traditional diagnostic methods for laryngeal pathologies, such as stroboscopy or manual analysis of acoustic signals, are time-intensive, operator-dependent, and often limited in sensitivity for early-stage or complex cases. The development of a non-invasive, objective and reliable laryngeal pathology detection and classification system could have a significant impact on the speed and accuracy of a delivered diagnosis.
- Challenges of Existing Automated Systems. The research on laryngeal bioimpedance as a single modality in pathology detection reports mixed results, with some authors suggesting EGG alone is unreliable (Borsky *et al.*, 2017; Nacci

et al., 2020; Miliaresi *et al.*, 2022; Tomaszewska and Georgakis, 2023). Furthermore, while most of research focuses on sustained phonation (Rosa *et al.*, 1999; Godino-Llorente and Gomez-Vilda, 2004; Henríquez *et al.*, 2009; Arias-Londoño *et al.*, 2010; Markaki and Stylianou, 2011; Hemmerling *et al.*, 2016; Al-Nasheri *et al.*, 2017; Harar *et al.*, 2017; Borsky *et al.*, 2017; Zhou *et al.*, 2022), there is a significant lack of literature reporting on the effectiveness of continuous speech signals in laryngeal pathology classification. This research seeks to investigate the classification performance of laryngeal bioimpedance (as well as audio) as a singular modality to find the most optimal methodology for the highest classification accuracy. The hypothesis is formed that the application of the ERB-derived feature extraction methods on continuous speech instead of sustained phonation improves the performance of pathology classification based on each investigated data modality.

- Potential of Multimodal Integration. Single-modality ("unimodal") approaches in automated pathology classification are limited in their capacity to capture the full range of complementary information inherent in both modalities. Combining audio and laryngeal bioimpedance signals offers a unique opportunity to leverage the strengths of each modality, providing a more holistic view of laryngeal behaviour. This research seeks to develop an accurate and efficient multimodal deep learning system that utilises audio signals and simultaneously recorded laryngeal bioimpedance, capable of detecting and classifying an existing laryngeal pathology.
- Data Availability and Quality Issues. There are very few publicly available datasets containing audio and simultaneously recorded laryngeal bioimpedance. Many existing databases suffer from limitations such as insufficient sample sizes,

inconsistent labelling, or emphasis on speaker-dependent features. This research seeks to establish a new database containing high quality audio recordings and simultaneous laryngeal bioimpedance collected from a control group of healthy participants, as well as those suffering from various laryngeal conditions, with a particular focus on cancerous lesions. Rigorous quality control and accurate labelling are intended to make this dataset a reliable resource for future research.

In response to the challenges listed above, in this study we develop a novel accurate and efficient multimodal deep learning framework that integrates features extracted from audio and EGG signals. For this purpose, multiple feature extraction and classification methods are investigated on both modalities, and the best performing models are chosen for both audio and laryngeal bioimpedance. By addressing the issue of the limited data through the collection and curation of a high-quality, synchronised dataset, this research provides a foundation for reliable machine learning applications. The study focuses on the classification of laryngeal pathologies across three key categories: healthy individuals, cancerous and precancerous growths, and neuromuscular disorders.

1.3. AIMS AND OBJECTIVES

The aim of this research is the development of a non-invasive, reliable and efficient laryngeal pathology detection and classification system, utilising audio signals and simultaneous bioimpedance measurements in a multimodal approach.

The following core objectives have been established:

1) DATASET: Development of a standalone dataset of combined audio signals and simultaneously recorded bioimpedance measurements. The data, including sustained phonation signals as well as continuous speech, has been collected from a control group of healthy participants, as well as those affected by the following

laryngeal disorders: malignant growths of the vocal fold area, other growths outside of the vocal fold area, vocal fold paralysis, laryngitis, Reinke's Oedema, and functional dysphonia. More information on the developed dataset and the classes used for the development of the final laryngeal pathology classification system can be found in chapter 4 of this thesis.

2) SPEECH vs SUSTAINED PHONATION: Comparison of laryngeal pathology classification performance based on continuous speech and sustained phonation. The feasibility of speech and sustained phonation for pathology detection has long been debated, with no clear consensus. This research provides a comparative analysis, demonstrating the advantages of speech signals in retaining diagnostically relevant features, thus advancing the understanding of their clinical utility. The results and partial conclusions on speech outperforming sustained phonation can be found in chapter 8 of this thesis.
3) FEATURE EXTRACTION: Selection of the most appropriate feature extraction algorithms for both data modalities – audio and bioimpedance. To optimise the performance of the final laryngeal pathology classifier, various feature extraction techniques are systematically investigated, with a particular focus on methods related to the ERB spectrum. We hypothesise that the ERB-derived features deliver the best performance for both data modalities, as they reflect critical spectral and physiological characteristics relevant to pathological signal analysis. The results and partial conclusions on the ERB-derived features outperforming other feature representations can be found in chapter 8 and 9 of this thesis.
4) PATHOLOGY DETECTION: The development of a reliable and accurate laryngeal pathology detection system. Two unimodal binary classification models differentiating between control and pathological cases are developed; one based on audio signals,

and the second utilising laryngeal bioimpedance measurements. Furthermore, a multimodal binary classification models differentiating between control and pathological cases is developed. The results are further described in chapter 8 and 9 of this thesis.

5) PATHOLOGY CLASSIFICATION: The development of a reliable and accurate laryngeal pathology classification system. For each modality – speech and sustained phonation – a separate unimodal classification system is developed, capable of differentiating between three major groups of signals: healthy, neuromuscular, and cancerous and precancerous growths (“malignant”). The obtained results are discussed in chapter 8.
6) MULTIMODALITY: Development of a final multimodal laryngeal pathology detection and classification system. The intended system relies on a multimodal approach to incorporate both types of data – audio and laryngeal bioimpedance collected using EGG. Three fusion approaches are investigated and compared: early fusion, hybrid (intermediate) fusion, and late fusion. Utilising the best-performing models for each modality, the final multimodal laryngeal pathology classification system is developed. The multimodal framework integrates complementary features from both modalities, enhancing diagnostic accuracy and establishing new benchmarks for pathology detection. The results produced by the multimodal system designed for the purposes of this study can be found in chapter 9.

1.4. CONTRIBUTIONS TO KNOWLEDGE

This research has led to several novel contributions to the field of laryngeal pathology detection and classification through the integration of audio and bioimpedance (electroglottographic) signals. These contributions are summarised as follows:

1. **Development of a standalone multimodal dataset (Chapter 4)**.

A unique dataset was developed, comprising audio recordings and simultaneous laryngeal bioimpedance measurements from both a control group of healthy participants, as well as individuals with a range of laryngeal pathologies. The dataset focuses particularly on precancerous and cancerous lesions, providing a valuable resource for future research in pathology classification. The publication of the developed dataset, alongside its statistical analysis, is currently in preparation.

2. **Comprehensive analysis of the collected data (Chapter 4)**.

The collected audio and bioimpedance signals were analysed using statistical, time-domain, and frequency-domain techniques – this was done to assess their fundamental characteristics and potential for separability. Clustering tendencies and grouping behaviours within the data were explored, providing initial evidence of class-specific features. Principal Component Analysis (PCA) was applied to visualise and quantify the separability of classes, uncovering patterns and correlations that informed subsequent classification processes. This analysis is intended to be published alongside the dataset.

3. **Systematic evaluation of feature extraction methods for laryngeal pathology detection and classification (Chapter 8)**.

A comprehensive comparative analysis of feature extraction methods for both audio and bioimpedance data modalities was conducted, with the identification of optimal techniques for each modality. Notably, features derived from the Equivalent Rectangular Bandwidth (ERB) spectrum demonstrated significantly superior performance compared to alternative

methods, including the Mel spectrum, which has been a longstanding benchmark for phonation analysis.

4. **Comparative analysis of the feasibility of continuous speech over sustained phonation for laryngeal pathology detection and classification (Chapter 8)**.

Contrary to prevailing literature and existing laryngeal pathology classification systems (Rosa *et al.*, 1999; Henríquez *et al.*, 2009; Markaki and Stylianou, 2011; Islam *et al.*, 2022), this work provides compelling evidence that continuous speech significantly outperforms sustained phonation in machine learning and deep learning-based systems for laryngeal pathology detection and classification.

5. **Systematisation of knowledge on electroglottography in medical applications for laryngeal health (Chapter 3)**.

During the course of this research, we produced and published a journal paper that systematises currently existing knowledge on electroglottography and the use of laryngeal bioimpedance in the medical diagnostics (Tomaszewska, J.Z. and Georgakis, A. (2023) 'Systematic Review of Electroglottography in Diagnostics, with Emphasis on Its Implementation in Digital Vocal Tract Pathology Classification Systems'. *Journal of Voice*, December 2023).

6. **Identification of data-related limitations within the existing research (Chapter 4)**.

Through the experimentation completed in this work we highlighted the significant limitations in existing laryngeal pathology datasets, particularly regarding speaker-dependent biases, data volume, and representation of specific conditions, such as

cancerous and precancerous lesions. More importantly, we proved the negative influence of speaker-dependent bias present in publicly available datasets on machine learning-based laryngeal pathology classifiers by conducting classification experiments with and without the application of a custom-built participant shuffling algorithm. Speaker-dependent biases constitute a “research trap” leading to artificially inflated results, thus, affecting all research resulting from such datasets. Such a speaker-dependent bias was identified within the Saarbruecken Voice Dataset (SVD), with possible broader implications for studies relying on this resource. We therefore review this popular database and make recommendation on its correct use.

7. **Development of an accurate and unbiased laryngeal pathology detection system (Chapter 9)**.

Two accurate unimodal systems were developed for laryngeal pathology detection – binary classification between healthy and pathological signals – one based on audio signals and the other on laryngeal bioimpedance measurements (Chapter 8). The third model based on the multimodal approach combining both data modalities was also proposed; with the best performance, the multimodal pathology detection system validated the hypothesis that multimodality facilitates more accurate detection of laryngeal pathologies compared to single-modality models (Chapter 9). All three systems address speaker-dependent bias and achieve robust classification performance.

8. **Development of an accurate and unbiased laryngeal pathology classification system (Chapter 9)**.

A novel multimodal classification system combining audio and laryngeal bioimpedance data was developed, with a particular focus on detecting precancerous and cancerous

lesions. This system represents a significant advancement in unbiased and accurate laryngeal pathology classification, leveraging the complementary strengths of both data modalities. To choose the best performing model for both data modalities, two accurate unimodal laryngeal classification systems were developed; one relying on audio, and the other on laryngeal bioimpedance.

9. **Systematisation of knowledge on multimodal fusion approaches in classification of laryngeal pathologies (Chapter 9)**.

A comprehensive comparative analysis of multimodal fusion strategies for laryngeal pathology classification was conducted, including early, hybrid (intermediate), and late fusion approaches. We identify the most effective fusion strategy for combining audio and laryngeal bioimpedance data for both detection of laryngeal pathologies, as well as the classification of laryngeal conditions with the focus on cancerous and precancerous lesions. This work advances the understanding of how different multimodal fusion techniques can enhance classification performance by leveraging the complementary strengths of each modality.

1.5. SUMMARY OF PUBLICATIONS

Tomaszewska, J.Z., Chousidis, C. and Donati, E. (2022) 'Sound-Based Cough Detection System using Convolutional Neural Network'. *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (pp. 1-6). IEEE, 2022, June.

Direct contribution to knowledge and research relevance: Assessment of CNN as an appropriate classification system for audio signals gathered from participants affected by laryngeal pathologies.

This publication presents a system for detecting coughs using audio signals analysed through a Convolutional Neural Network (CNN). Although it primarily focuses on respiratory sounds, the methodologies developed – particularly in feature extraction and deep learning for audio classification – are directly applicable to the research on audio recordings of patients and their application in laryngeal pathology detection and classification. This paper showcases that the chosen feature extraction and deep learning architecture (Convolutional Neural Networks) are appropriate for the analysis of the audio signals obtained from patients affected by laryngeal pathologies.

Tomaszewska, J.Z., Młyńczak, M., Georgakis, A., Chousidis, C., Ładogórska, M. and Kukwa, W. (2023) 'Automatic Heart Rate Detection during Sleep Using Tracheal Audio Recordings from Wireless Acoustic Sensor'. *Diagnostics*, 13(18), p.2914.

Direct contribution to knowledge and research relevance: Assessment of the potential of audio signals in physical health assessment.

This paper contributes as the preliminary work aimed at assessing potential of audio recordings gathered from the upper part of the body in detection of pathologies within that area. This work demonstrates the use of audio signals for physiological monitoring. While the focus is on cardiovascular monitoring, it showcases the potential of audio signals as a source of information for appropriate physiological assessment, hence contributing as potential diagnostic tool.

Tomaszewska, J.Z. and Georgakis, A. (2023) 'Systematic Review of Electroglottography in Diagnostics, with Emphasis on Its Implementation in Digital Vocal Tract Pathology Classification Systems'. *Journal of Voice*, December 2023.

Direct contribution to knowledge and research relevance: Systematisation of available literature on EGG in laryngeal pathology diagnostics – gathering academic evidence for the hypothesis of electroglottography being an appropriate tool for laryngeal pathology detection.

This systematic review was written to provide an in-depth analysis of the past and the current use of electroglottography (EGG) in laryngeal pathology diagnostics. This paper is crucial to this research, as it allowed us to consolidate the existing knowledge identifying the gaps, and to assess the theoretical capabilities (potential) of electroglottographic signals as a medium for laryngeal pathology detection. Based on this review, we put forward the hypothesis of electroglottography contributing as a valuable tool for non-invasive laryngeal pathology detection. This helped forming the foundation for combining EGG with audio signals in a multimodal classification system.

Tomaszewska, J.Z., Chousidis, C. and Georgakis, A. (2024) 'Comparative Analysis of MFCC and GTCC Performance in Laryngeal Pathology Detection Based on Electroglottographic Signals'. *Proceedings of the Institute of Acoustics 2024* (Vol. 46. Pt. 2. 2024). Institute of Acoustics, 2024, September. DOI: 10.25144/23671

Direct contribution to knowledge and research relevance: Evidencing that the ERB-derived features (such as GTCC evaluated in this paper) outperform Mel spectrum-derived features (in this case, the MFCC) in laryngeal pathology detection systems. Also, evidencing that speech signals provide more information for the detection of laryngeal pathology than sustained phonation – contrary to most currently available literature.

This paper presents a comparative analysis of two feature extraction methods, Mel-frequency Cepstral Coefficients (MFCC – derived from the Mel Spectrum) and Gammatone Cepstral Coefficients (GTCC – derived from the Equivalent Rectangular Bandwidth

Spectrum), for detecting laryngeal pathologies using laryngeal bioimpedance (electoglottographic) signals. The two types of phonation – continuous speech and sustained phonation – were compared in their ability to preserve features relevant for laryngeal pathology detection. The study finds that GTCCs outperform MFCCs in capturing relevant features from EGG signals, offering enhanced sensitivity for detecting laryngeal pathologies. Furthermore, it was found that the electroglottographic speech signals retain more information relevant for pathology detection than electroglottographic measurements of sustained phonation – contrary to most currently available literature. This work has been awarded “Best Poster” during the ACOUSTICS 2024 Conference.

Donati, E., **Tomaszewska, J.Z.** and Chousidis, C. (2024) ‘Evaluation of $f0$ Stability in Speech and Singing Using Laryngeal Bioimpedance Measurements’. *Proceedings of the Institute of Acoustics 2024* (Vol. 46. Pt. 2. 2024). Institute of Acoustics, 2024, September. DOI: 10.25144/23655.

Direct contribution to knowledge and research relevance: Assessment of the variation of fundamental frequency ($f0$) in speech and sustained phonation, evidencing both phonation types can be distinguished upon statistical differentiation.

This study establishes a clear statistical differentiation between speech and sustained phonation based on the variability of their $f0$. Utilising laryngeal bioimpedance (electroglottographic measurements), the study demonstrates that speech exhibits significantly higher $f0$ variability compared to sustained phonation, providing an objective basis for distinguishing these vocal modes. These findings are particularly relevant to our research on laryngeal pathology classification, as they align with observed performance trends where classification systems achieve superior accuracy with speech over sustained phonation.

1.6. THESIS STRUCTURE

This thesis consists of ten chapters that encompass the work conducted in analysing the literature and achieving the stated aim and objectives. This section provides an overview of the thesis structure and its contents.

Chapter 1 (*Introduction*) introduces the subject of laryngeal pathology, their diagnostics, and significance for public health. It provides an initial overview of the current state-of-the-art techniques and methodologies in laryngeal pathology classification, identifying gaps and challenges in the existing research. Subsequently, the chapter outlines the aims and objectives of the research, stating the research rationale and the research hypothesis. A comprehensive list of novel contributions to knowledge is provided, followed by a detailed record of the publications produced during the course of the research, demonstrating the scholarly impact of the work.

Chapter 2 (*Background Theory on Human Phonation and Laryngeal Pathologies*) lays the key background knowledge relevant for laryngeal pathology detection and classification implemented in this study; this includes the introduction to human phonatory apparatus and the nature of human-made sounds in a recorded digital form, comprising audio and laryngeal bioimpedance signals. Lastly, all specific laryngeal pathologies discussed in this thesis are introduced from the medical and analytical perspective; among others, those conditions include precancerous and cancerous lesions, neuromuscular disorders, and others.

Chapter 3 (*Review of Literature on Laryngeal Pathology Classification Systems*) critically examines the state of the art in laryngeal pathology classification, focusing on the use of audio and laryngeal bioimpedance signals (electroglottography). It provides the analysis of the existing digital systems of pathological signal classification, highlighting the strengths and limitations of explored methodologies, and identifying the areas for improvement. Similarly, it reviews the application of audio signals and laryngeal bioimpedance

measurements in the field, identifying gaps in research and emphasising the need for more robust approach to feature extraction methods.

Chapter 4 (*Datasets and Data Initial Analysis*) describes the data used in this research. Details are provided on data collection, participant demographics, and preprocessing techniques. The chapter also provides a thorough data analysis and explores the class separability through statistical and spectral analyses, using PCA, Hopkins statistics and Euclidean Distance measures to investigate data's clustering tendencies. Furthermore, the initial class setup and the final selection of the dataset chosen for the research are explained. Lastly, the use of Saarbruecken Voice Database is discussed, providing a list of its limitations and an overview of preprocessing methods used to mitigate them.

Chapter 5 (*Feature Extraction Methods*) discusses feature extraction techniques used for both audio and bioimpedance data. It compares time-domain and frequency-domain representations, evaluating their effectiveness in capturing diagnostic information. Key methods, including the Equivalent Rectangular Bandwidth (ERB) spectrum, and Mel spectrum are explored in detail. Each feature extraction method is analysed, with the results of the classification based on each method presented, investigated and described in detail. In this chapter we argue that the ERB-derived features have the potential to outperform traditional Mel-based methods, providing insights into their physiological relevance.

Chapter 6 (*Machine Learning and Deep Learning Methods*) introduces the classification algorithms used for discrimination between laryngeal pathologies. It outlines the theory behind each classification method, as well as the architecture chosen for each investigated model; those include Random Forest, CNNs (one-dimensional and two-dimensional), and recurrent architectures such as LSTMs and BiLSTMs. The machine learning and deep learning methods are explained.

Chapter 7 (*Multimodality*) introduces the concept of multimodality for laryngeal pathology classification, combining audio and bioimpedance data. It discusses three fusion strategies: early fusion, hybrid fusion, and late fusion, including the stacked generalisation, analysing the details of each approach and highlighting their respective advantages and challenges.

Chapters 8 (*Unimodal System Results*) and 9 (*Multimodal System Results*) present the results of: the unimodal laryngeal pathology detection, the unimodal laryngeal pathology classification focused on detection of cancerous and precancerous lesions (chapter 8), and the multimodal detection and classification (chapter 9). In chapter 9 the results for each fusion strategy are presented and evaluated for both detection and classification of the laryngeal pathologies. This chapter further consolidates and directly compares the results of the various classification systems and feature extraction methods explored throughout the research. It provides the discussion of the findings, contextualising them within the broader field of laryngeal pathology diagnostics. In chapter 8 we provide the results of statistical significance testing (one-way ANOVA and Tukey's HSD) of continuous speech versus sustained phonation comparison for multi-class laryngeal pathology classification with detection of cancerous and precancerous lesions, while in chapter 9 the statistical significance testing is broadened to cover the comparison of all multimodal systems and best-performing unimodal systems.

Finally, chapter 10 (*Conclusions*) concludes the thesis by summarising the key findings and their significance. The hypothesis is restated, and the contributions of this research to the field are reaffirmed. The chapter also outlines potential avenues for future research, including suggestions for improving datasets and extending the use of the developed systems to a real-time application and real-life medical diagnostics.

All references used throughout the study can be found in chapter 11. The appendices, including tables of detailed result parameters, as well as the summary table containing the history of electroglottography, can be found in chapter 12.

Additionally, we provide a folder of the algorithms written for the purposes of this research which is available upon request.

Chapter 2

Background Theory on Human Phonation and Laryngeal Pathologies

To understand laryngeal pathologies and their classification, it is crucial to acknowledge the anatomy of the human phonatory system. The direct source of sound produced during speech or singing are the vocal folds – vibrating structures, that reside within the larynx. Any abnormalities caused to those structures can alternate their vibration pattern, possibly affecting the laryngeal bioimpedance readings, as well as the acoustic properties of the voice itself. To record and detect the signal changes indicating a plausible presence of a laryngeal pathology, it is essential to fully comprehend the human phonation process, as well as the technicalities of both data modalities.

The following chapter discusses the anatomy of the human phonatory system and the physiological mechanisms critical for sound production. Subsequently, the two primary data modalities used to analyse phonation are discussed in context of human phonation and its representation in the digital domain: the audio signals, capturing acoustic characteristics, and laryngeal bioimpedance, offering insights into underlying physiological processes. Finally, the concept of laryngeal pathologies is explored, providing description and examples of each laryngeal abnormality investigated in this research. In this section, we also present the figures of audio and bioimpedance waveforms and their frequency spectrum representations that were obtained from the signals of the custom dataset developed for the purposes of this study (see chapter 4).

2. BACKGROUND THEORY ON HUMAN PHONATION AND LARYNGEAL PATHOLOGIES

2.1. HUMAN PHONATORY SYSTEM

The human phonatory apparatus is directly linked to the respiratory system, as phonation relies on the airflow generated during respiration. The respiratory system is therefore integral to two fundamental functions – breathing, as well as voice generating (Moussavi, 2006). It comprises the diaphragm, the rib cage, lungs, the trachea-bronchial tract, the larynx, and upper airways that include pharynx, nasal cavities, and oral cavities (Marchal, 2009). The main purpose of the human phonation apparatus is the conversion of kinetic energy (specifically, aerodynamic energy) delivered from the flow of air from lungs during the respiratory cycle into the acoustic energy (Rossing, 2007). Both the breathing, as well as the voice generating processes rely on the constant flow of air throughout the respiratory cycle, which consists of two phases: inhalation and exhalation (Marchal, 2009). During the inhalation, the diaphragm contracts and the rib cage expands, allowing the air to travel from nasal or oral cavities to the pharynx. Then, the air reaches the larynx (a simplified depiction of a human larynx can be seen on Figure 2.2), where the vocal folds (also referred to as vocal cords) reside (Rossing, 2007). Finally, through the trachea and bronchial tree, the air finds its way into the lungs (Moussavi, 2006). After the gas exchange of oxygen and carbon dioxide in blood cells, the exhalation process begins – the diaphragm relaxes, emptying the lungs and expelling gaseous waste from the body.

Figure 2.1 depicts a simplified model of the human respiratory system.

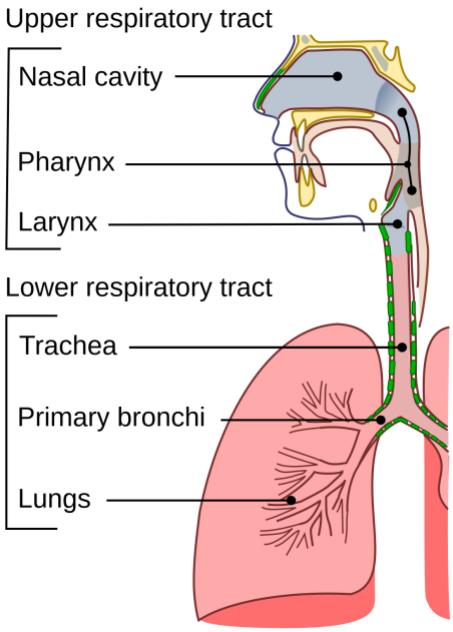

Figure 2.1: Model of human respiratory system (image directly sourced and reproduced from Wikimedia Commons (2010)).

The role of human phonatory apparatus becomes most prominent during the exhalation process. Most speech related sounds are produced as the airflow passes through the glottis – the space in between the vocal folds (Marchal, 2009). The folds – abducted (open) during breathing – assume the adducted (closed) position as the process of phonation proceeds. The steady stream of airflow generated during the exhalation process places the vocal cords in a state of self-sustained oscillation as the vibratory tissue motion produces an acoustic wave (Redford, 2019), thereby acting as a sound generator. This is possible due to Bernoulli's principle, according to which an increase in the velocity of airflow results in a decrease in pressure within the flow (Marchal, 2009). When air is expelled from the lungs and passes through the glottis, the airflow velocity increases due to the narrowing of the glottal space. This creates a pressure drop between the vocal folds, causing them to be drawn together. Once the folds are momentarily closed, subglottal pressure builds beneath them until it forces them apart, initiating a new vibratory cycle. The oscillation proceeds as the paired symmetrical lateral cricoarytenoid and interarytenoid muscles pull the vocal folds

together. The constant airflow creates the subglottal pressure as the air passes through the glottis. This high-speed air creates a suction effect and brings the vocal folds together (Marchal, 2009). This recurring process gives rise to variations in pressure levels that characterise the resulting sound.

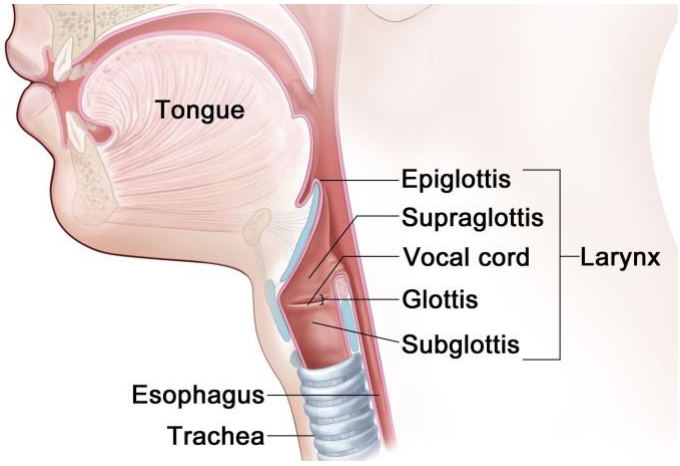

Figure 2.2: Model of human larynx (image directly sourced and reproduced from www.teresewinslow.com (2012)).

In the final part of the process, the air reaches the pharyngeal cavity (pharynx), mouth cavity, and nasal cavity (Netter, 2019). Those elements of the human respiratory system act as resonators, shaping the sounds generated by the vocal cords (Sataloff, 2017). The vocal tract can therefore be thought of as a distinctive filter that shapes a simple acoustic wave generated within the larynx into a distinguishable human voice, unique for each specimen.

Considering the crucial role of the larynx in the phonation process, it is reasonable to state the vocal folds are the direct source of human voice. The folds' oscillation and the change in the larynx position cause a variety of bioimpedance alternations that could be immensely informative during the phonatory system diagnostic process. The application of EGG allows for the continuous and direct observation of those parameters. Nevertheless, all resonant cavities of the pharynx shape the produced sound into a distinguishable voice. Since the cavity of the larynx also acts as the resonator, any impairment within that area that does not necessarily affect the vocal folds directly could result in voice changes, with

the bioimpedance remaining unchanged. This leads to a conclusion that while EGG can capture vital changes in bioimpedance related to laryngeal pathologies, the combined assessment of bioimpedance and audio signals could produce more accurate and reliable assessment of vocal tract health status.

2.2. HUMAN PHONATION – AUDIO SIGNALS

Auditory evaluation is one of the primary methods for examining the respiratory and laryngeal function and the vocal tract health status. There is a vast amount of research on sounds produced within the human respiratory system being crucial for the correct identification of an illness and its required treatment (Sarkar *et al.*, 2015; Rao *et al.*, 2018). This applies in conditions related to the lower respiratory system, such as various pulmonary diseases, as well as the assessment of upper respiratory conditions – for instance, disorders affecting the larynx.

Nevertheless, the limitations and subjectivity of human hearing provides a threshold in the efficiency of this method (Sarkar *et al.*, 2015). A foreseen alternative for voice auscultation performed by a clinician, that could provide higher accuracy and objectivity, is the development of a digital voice assessment system of audio recordings gathered from affected participants. To fully comprehend this concept, a thorough understanding of audio as a data source is necessary.

An audio signal is an electrical representation of a sound in a form of a voltage or current fluctuation (Davis *et al.*, 1989). The sound level, expressed in decibels (dB) and analogous to the signal's amplitude, is directly related to the amplitude of the passing voltage. Beside the amplitude, this signal encapsulates other critical characteristics of sound, including its frequency, and phase. Frequency, measured in Hertz (Hz), corresponds to the pitch of the sound, with lower frequencies representing bass notes and higher frequencies for treble

notes. The phase of an audio signal reveals the temporal alignment of the sound wave within its periodic cycle, enabling the precise reconstruction of intricate acoustic phenomena. The audio signals serve as the fundamental carriers of auditory data and are vital in a myriad of applications, including music production, telecommunications, and multimedia transmission, as well as speech recognition, bioacoustics, and a wide range of bio-medical research.

An audio signal produced by human phonatory system can be referred to as the voice. Voice generation begins in the lower respiratory system, where the constant flow of air is produced during the exhalation process. Then, the generated aerodynamic energy induces the self-sustained oscillation of the vocal cords within the larynx. From there, the generated sound travels through the pharynx, mouth cavity, and nasal cavity which serve as resonators, shaping and modifying the sound produced by the vocal folds. Any alterations to the shape and size of these resonators produce changes in speech sounds known as phonemes, which are key to the formation of languages. This leads to a conclusion that any pathological changes within resonance cavities may alter the sound of human voice, which confirms the theory that auditory voice assessment is one of the key elements in phonatory tract diagnostic processes.

In the digital domain, an audio signal is a discrete representation of sound waves, encoding oscillations in air pressure as sequences of numerical values. This digital representation preserves characteristics such as amplitude (related to loudness), frequency (perceived as pitch), and phase (indicating temporal alignment). Furthermore, the ability to process audio signals digitally enables large-scale data analysis, cross-validation of findings, and the possible development of predictive models. By capturing sound properties, computational methods can aid to analyse the signal in depth, extracting features that correlate with physiological changes in the vocal mechanism.

The utilisation of audio signals holds immense potential in the fields of vocal tract auscultation and laryngeal pathology detection. Considering the anatomy of a human phonatory apparatus and resonant capabilities of the larynx and pharynx cavities, human produced sounds can offer invaluable insights into the functioning of the vocal tract. The analysis of various acoustic properties, such as fundamental frequency ($f0$), signal-to-noise ratio (SNR), harmonic-to-noise ratio (HNR), or even autocorrelation features, as well as the spectral characteristics including spectral entropy, spectral centroid, or spectral spread, may allow for the identification of structural irregularities in the phonatory tract and enable the assessment of the efficiency of vocal fold vibration. Additionally, advances in machine learning and signal processing techniques enable a more thorough analysis of the features derived from the signals gathered from participants affected by any laryngeal abnormalities, possibly leading to a development of a system for assessing laryngeal pathology that could contribute towards a more accurate and efficient healthcare diagnostics.

In summary, the rich spectral content of an audio signal influenced by the anatomical build of the phonatory tract, and thereby any impairment found within, may be highly relevant for the laryngeal pathology diagnostics. Furthermore, the non-invasive nature of audio signal collection and assessment holds significant promise for the development of a reliable diagnostic tool.

2.3. HUMAN PHONATION – BIOIMPEDANCE SIGNALS (ELECTROGLOTTOGRAPHY)

To understand the role of electroglottographic (EGG) measurements in the detection and classification of laryngeal pathologies, it is first necessary to fully comprehend the principles of bioimpedance. In electronics, the measurement of impedance quantifies the opposition to alternating current (AC), accounting for both resistance and reactance in a circuit. It is mathematically expressed as:

$$Z = \sqrt{R^2 + (X_L - X_C)^2} \quad (2.1)$$

Where Z represents the impedance, which is the root of the squared resistance (R^2) and the squared difference between inductive reactance (X_L) and capacitive reactance (X_C).

In biological tissues, impedance mirrors the interaction between the electrical properties of cells and alternating current. Intracellular and extracellular fluids contribute resistance, while the lipid bilayer of cell membranes introduces capacitance due to its insulating nature between conductive protein layers. Consequently, the tissue's response to alternating current is frequency dependent. At low frequencies, current predominantly flows through extracellular fluids, bypassing cellular structures. At high frequencies, current penetrates cell membranes, engaging intracellular components. This frequency-dependent behaviour underpins the concept of bioimpedance in human tissues (Donati, 2022).

Electroglottography (EGG) applies these principles to measure bioimpedance changes in the larynx, capturing physiological phenomena associated with vocal fold vibrations during human phonation. These impedance variations, recorded as time-varying signals, show the changes in voltage or current flow corresponding to the cyclic opening and closing of the vocal folds during speech (Childers and Larar, 1984).

The EGG signal is therefore often interpreted as a measurement of the vocal fold contact area (VFCA), reflecting the dynamic conductivity changes as the vocal folds move closer together or farther apart. When the folds come into contact, tissue conductivity increases and impedance decreases; when the folds separate, impedance increases. These measurements offer valuable insights into vocal fold biomechanics, aiding the analysis of phonatory behaviour and voice production.

In the literature, EGG signals are typically referred to as Lx waveforms, which represent the bioimpedance signal after basic preprocessing (Fourcin and Abberton, 1971; Lecluse, 1975; Baken, 1992). Raw EGG signals, often termed Gx, include contributions from neck

tissues and must be high-pass-filtered to isolate the component related to vocal fold activity (Titze, 1990). The resulting Lx signal is a more specific representation of the vocal fold dynamics.

The electroglottographic evaluation involves placing two electrodes on either side of the thyroid cartilage (Figure 2.3). One electrode delivers a high-frequency, low-amperage current (passes the voltage), while the other records the resulting signal. The produced waveform (Figure 2.4) correlates with the glottal cycle phases, capturing transitions in VFCA during phonation (Herbst, 2019).

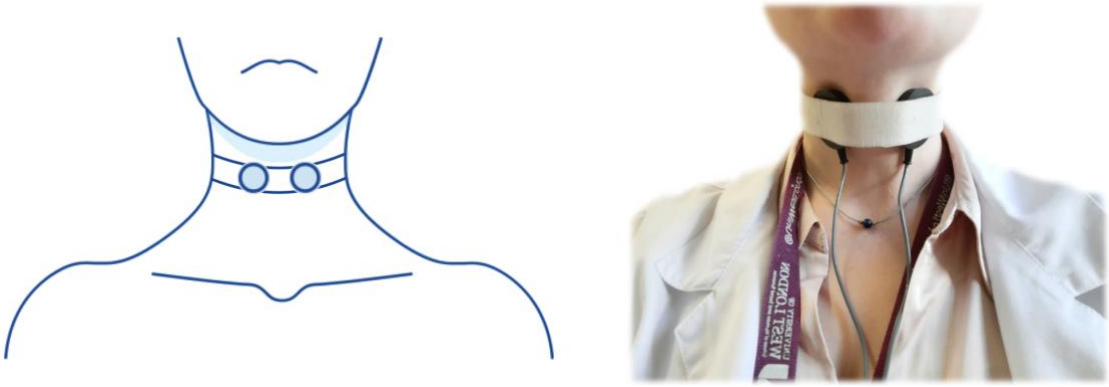

Figure 2.3: Electroglottography and its electrode placement.

The ideal stereotypical waveform of such signal can be observed in Figure 2.4, with all stages of glottal opening and closing described according to Childers *et al.* (1986; 1987), Rothenberg (1981), and Baken (1992).

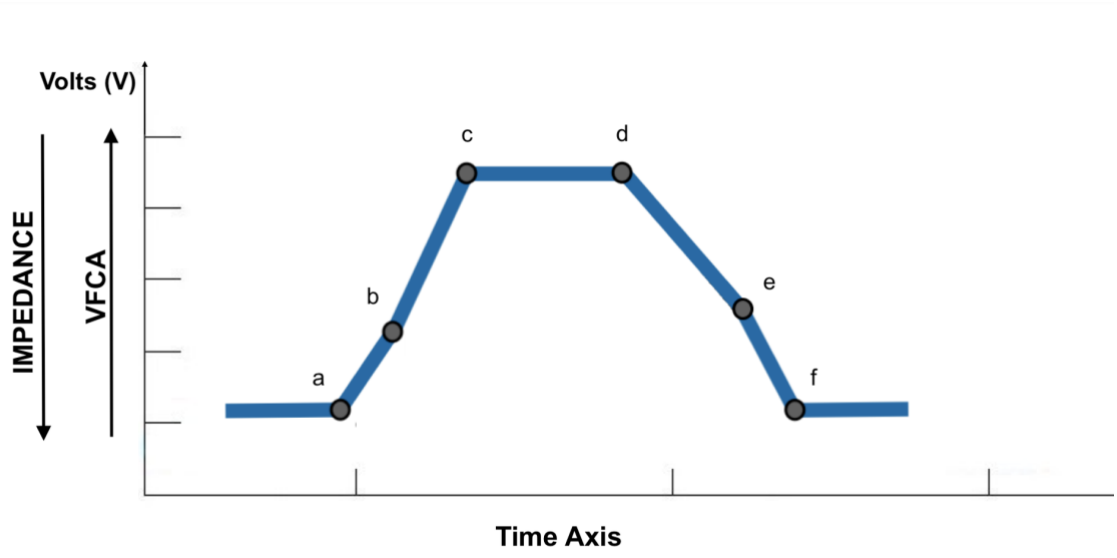

Figure 2.4: Idealised electroglottographic waveform, illustrating the relationship between the measured impedance, vocal fold contact area, and phases of the glottal cycle.

The stereotypical waveform illustrates the stages of the glottal cycle as follows (Herbst, 2019):

- a: Initial contact of the lower vocal fold margins (onset of the closing phase).
- b: Initial contact of the upper vocal fold margins.
- c: Maximum contact of the vocal folds (end of closing phase) – this glottal phase does not necessarily imply the actual complete contact of the vocal folds; instead, it shows the maximum contact for that particular glottal cycle.
- d: Initial separation of lower vocal fold margins (onset of the opening phase).
- e: Initial separation of upper vocal fold margins.
- f: Full glottal opening, with the minimal contact area between vocal folds.

In the existing body of literature, the visual representation of the electroglottographic (EGG) signal varies considerably, leading to confusion in interpreting the waveform (Baken, 1992). This inconsistency often stems from a lack of understanding of the specific electrical circuit configuration adopted in each study. Two predominant approaches for constructing EGG circuits include: (1) representing increasing signal amplitude as equivalent to

increasing impedance, and (2) associating increasing signal amplitude with increasing VFCA.

1) **Childers's representation** (Figure 2.5)**:** In this approach, the increase in the signal's amplitude corresponds to an increase in bioimpedance. This method has been utilised by researchers such as Childers *et al.* (1983; 1984; 1985; 1992), Colton and Conture (1990), and Rothenberg in his work on multichannel electroglottography (1992). Here, higher amplitudes signify decrease in VFCA.

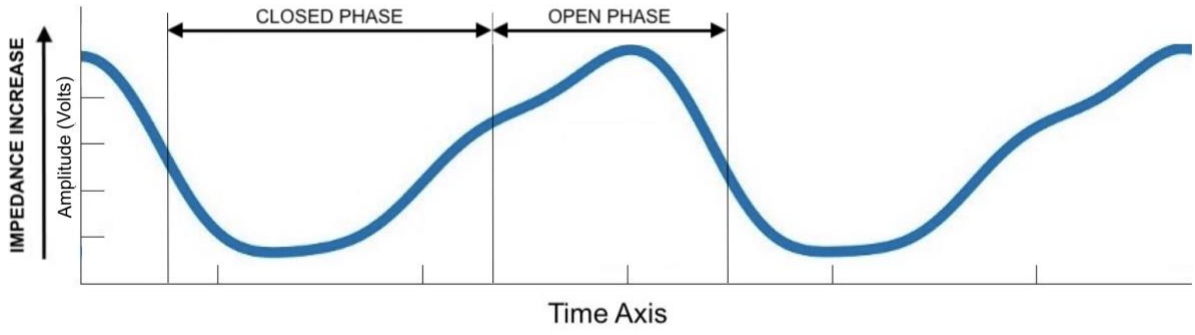

Figure 2.5: Childers's representation of EGG signal. Y-axis corresponds to bioimpedance; increasing amplitude indicates higher bioimpedance and reduced VFCA.

2) **Fourcin's representation** (Figure 2.6)**:** Introduced by the inventor of laryngograph himself, Fourcin (1971; 1972), this depiction is the most commonly used in recent literature (Deshpande and Manikandan, 2017; Baken, 1992; Herbst, 2019). In this representation, an increase in signal amplitude indicates a decrease in bioimpedance, corresponding to an increase in VFCA. This is also the form produced by widely used devices such as the Kay Pentax or Kay 6103 (Nacci, 2020).

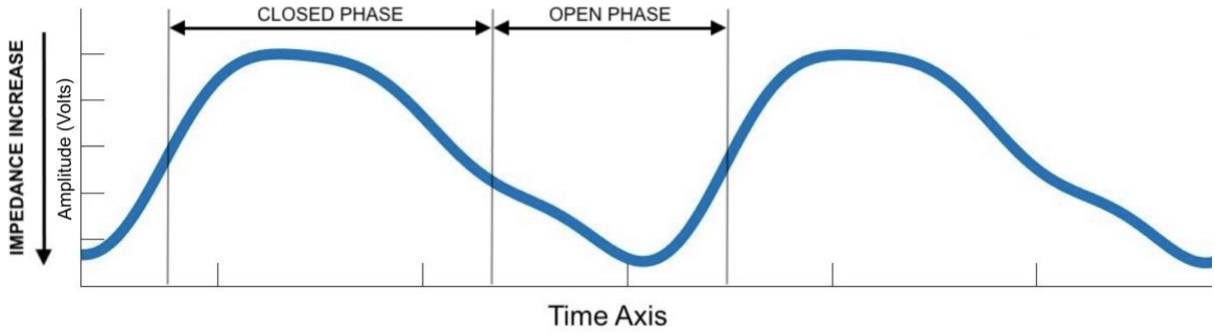

Figure 2.6: Fourcin's representation of EGG signal. Y-axis corresponds to VFCA; increasing amplitude indicates decreased bioimpedance and increased VFCA.

Recent literature, as well as commercially available electroglottographs, predominantly adopt the Fourcin approach, where signal amplitude increases in parallel to decreasing impedance. For consistency with modern standards and devices, this research also adopts Fourcin's representation, plotting EGG waveforms with VFCA on the Y-axis, where increased signal amplitude corresponds to reduced bioimpedance, and thus, the closed phase of phonation.

Given the close relation of laryngeal bioimpedance and VFCA, it can be expected for the electroglottographic signal to change depending on a presence of laryngeal impairment affecting the vocal folds, or the lack thereof. This leads to a hypothesis that the insights gained from laryngeal bioimpedance hold significant promise in the realm of laryngeal diagnostics.

The application of bioimpedance in laryngeal pathology detection and classification has profound implications, particularly when analysed in the context of computational methods and digital signal processing. Human phonation generates complex, time-varying signals that can be captured as bioimpedance measurements, reflecting the intricate physiological dynamics of the vocal folds. In a digital domain, these signals are transformed into discrete data points, enabling advanced computational analysis to extract meaningful patterns.

Digital processing techniques, such as time-frequency decomposition, spectral analysis, and feature extraction, facilitate deeper insights into vocal fold dynamics and provide a rich dataset for machine learning models. Such models can identify subtle changes, such as those caused by malignant growths or neuromuscular disorders, with high precision.

By leveraging the discrete nature of bioimpedance signals and integrating computational techniques, this research seeks to advance the understanding and application of EGG signals in the detection of laryngeal pathologies, offering significant potential for clinical and research applications.

2.4. INVESTIGATED LARYNGEAL PATHOLOGIES

For the purposes of this research, the conditions under investigation are grouped into three primary categories: cancerous and precancerous growths, neuromuscular disorders, and a control group of healthy participants. Neuromuscular disorders, such as vocal fold paralysis, arise from impairments in the nerves or muscles responsible for coordinating vocal fold movement. They are characterised by impaired mobility, often leading to glottic insufficiency and irregular phonation. These conditions present dynamic challenges in signal interpretation due to their time-varying and asymmetrical nature. In contrast, cancerous and precancerous growths, including laryngeal carcinomas and malignant tumours, involve the development of abnormal tissue that may obstruct regular phonation and pose serious health risks. Their altered vibratory patterns and biomechanical properties may be captured through bioimpedance and audio signal analysis. The healthy group serves as a baseline for comparison, providing insights into the normative ranges of vocal fold behaviour during phonation. This categorisation enables a structured analysis of pathologies with distinct mechanisms of onset and progression, while also facilitating the identification of signal characteristics unique to each group.

For the purposes of this research, a custom dataset was created containing the audio recordings and simultaneous laryngeal bioimpedance measurements (electroglottographic signals) collected from patients suffering with the following pathologies: malignant growths of the vocal fold area (including both cancerous and precancerous growths), other growths outside of the vocal fold area (including solely the benign growths), neuromuscular disorders (including predominantly vocal fold paralysis caused by an underlying damage to the recurrent laryngeal nerve), laryngitis, Reinke's Oedema, and functional dysphonia. Although all pathological data – data from all six groups of laryngeal disorders – was used for the development of the final binary laryngeal pathology detection system, not all the pathological subgroups were used in the development of the envisaged multi-class laryngeal pathology classification system. This was due to factors such as diagnostic inaccuracies, excessive intra-class variability, or the true classification of specific conditions as symptoms rather than distinct diseases. The excluded pathologies encompassed functional dysphonia, laryngitis, growths outside of the vocal fold area, and Reinke's Oedema. The exact reasons for the groups' rejection are further explained in the following subsections.

The choice of the remaining pathologies was based on two primary criteria: (1) they specifically impact the larynx and the vocal folds within it, resulting from physical or physiological impairments in this region, (2) they constitute to most prevalent laryngeal pathologies amongst affected patients. Such conditions are of high clinical relevance given their potential for serious health outcomes. For instance, approximately 60% of laryngeal cancers originate from growths on the vocal folds or within the glottis, while 35% begin in the supraglottic region (American Cancer Society, 2023). Early detection and diagnosis of these growths can be instrumental in preventing malignancies from progressing to advanced stages. Furthermore, since all chosen pathologies affect the larynx, our hypothesis is that

the resulting disruptions in vocal fold behaviour should be distinctly observable in both audio and electroglottographic signals.

In this section of the report, we list all pathologies investigated in this research and provide a short description of each disorder. For each investigated pathology, a figure representing a randomly selected sample from each category of the collected data is presented in a form of an audio signal, its bioimpedance counterpart, as well as the representation of their frequency content for each signal. The following figure (Figure 2.7) depicts the audio signal waveform alongside its bioimpedance counterpart, as well as the frequency spectrum representations for each data modality collected from a random participant from the control group of the custom dataset developed for the purposes of this study (see chapter 4).

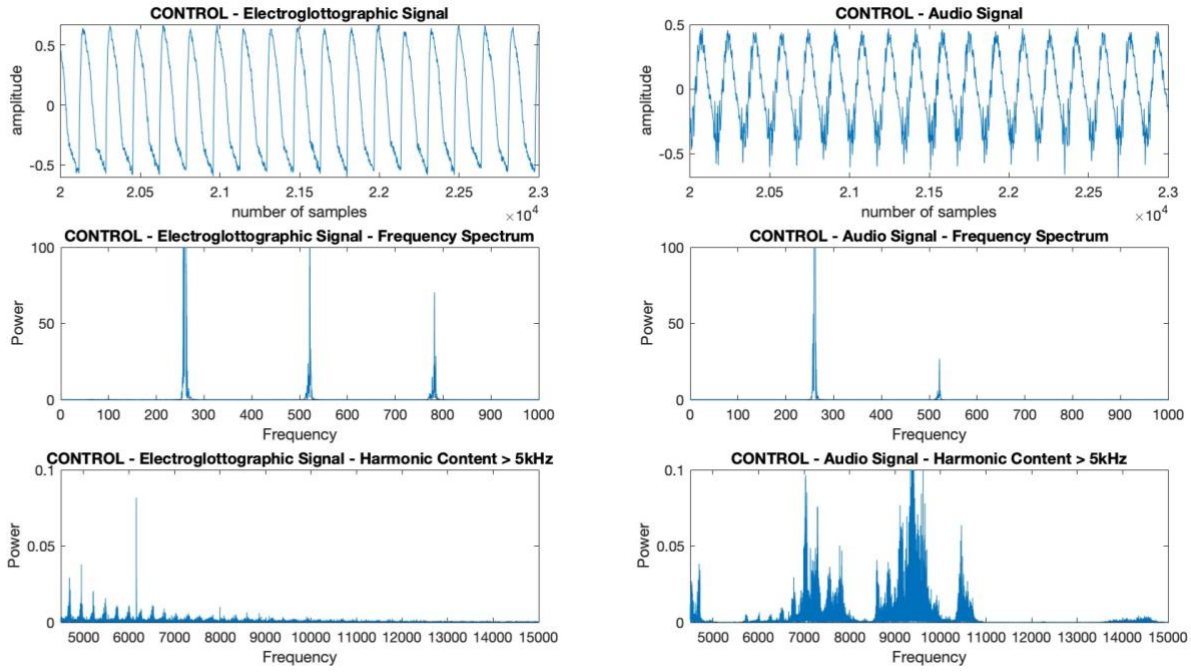

Figure 2.7: "Control signal". Representation of control signal – unaffected by any of the investigated pathologies – EGG signal from an individual from the control group (top left) and its spectral representation (bottom left), and audio signal from an individual.

2.4.1 Malignant Growths of Vocal Folds

The malignant growths of the vocal folds represent some of the most severe laryngeal pathologies investigated in this research, encompassing both cancerous and precancerous conditions. These growths often originate from cellular abnormalities within the glottis or supraglottic regions, accounting for the majority of laryngeal cancers; according to the American Cancer Society (2023), approximately 60% of laryngeal cancers arise from the glottis, while 35% develop in the supraglottic area. The precancerous lesions, such as leukoplakia, frequently precede malignant transformations and are often attributed to risk factors like prolonged exposure to tobacco smoke, alcohol consumption, or environmental carcinogens (Jones *et al.*, 2016).

These malignancies significantly disrupt the normal function of the vocal folds, leading to symptoms such as persistent hoarseness, difficulty breathing, and dysphonia. In advanced stages, tumours may invade surrounding tissues, causing pain and impairing swallowing (dysphagia). The presence of malignant growths often leads to alterations in the vibratory properties of the vocal folds, which is reflected in both audio and electroglottographic signals. Timely diagnosis of such conditions is critical, as early detection and treatment can significantly improve survival rates. Standard treatment approaches for malignant vocal fold growths typically include a combination of surgery, radiation therapy, and chemotherapy, depending on the stage and extent of the disease (Jones *et al.*, 2016). Advances in voice-preserving surgical techniques have further enhanced outcomes, allowing for improved quality of life in affected individuals.

In this study, the class of “malignant growths” includes both cancerous as well as precancerous lesions. This decision was influenced by both data limitations and clinical considerations. The separation of cancerous from precancerous lesions requires extensive medical investigation, often involving invasive procedures and specialist interpretation,

which were beyond the scope of this project; in this study, the priority was to detect the cancerous growths, as well as precancerous to enable the possible timely treatment in either case. Future work, with larger datasets and closer clinical collaboration, may enable a more fine-grained categorisation. For the current study, however, combining the two categories provided a clinically meaningful grouping while maintaining methodological robustness.

The following figure (Figure 2.8) depicts the audio signal waveform alongside its bioimpedance counterpart, as well as the frequency spectrum representations for audio signal and simultaneously recorded laryngeal bioimpedance collected from a participant suffering with a malignant vocal fold lesion.

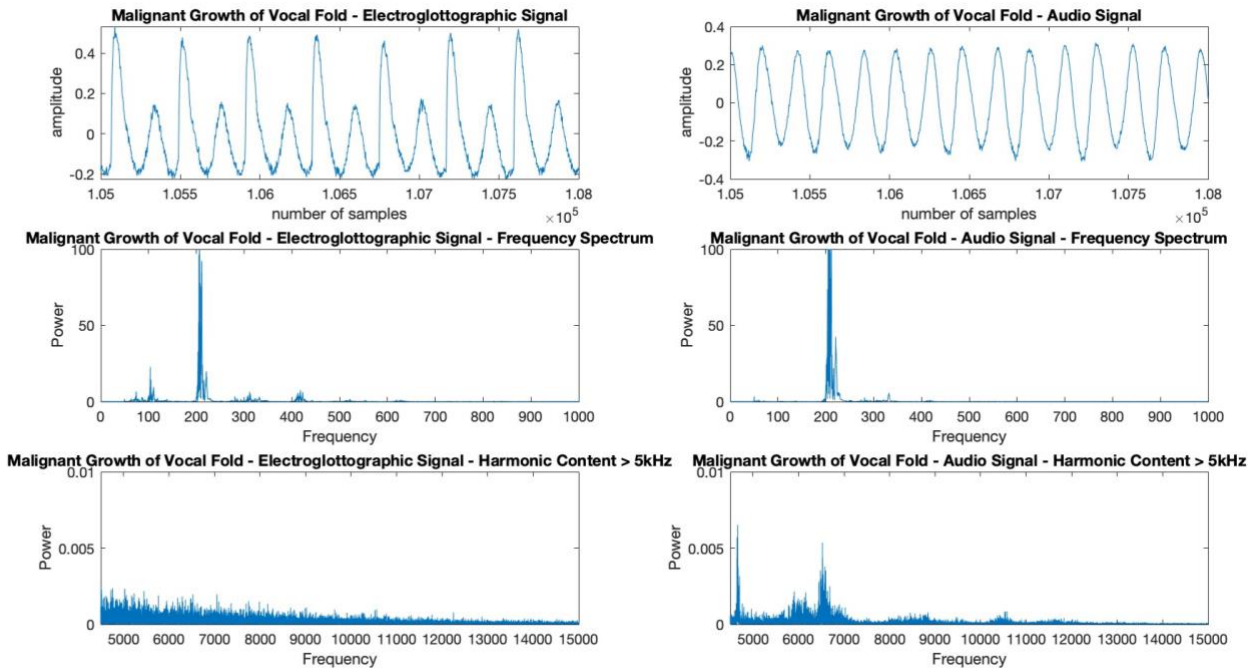

Figure 2.8: "Malignant Growth of Vocal Fold". Representation of a signal obtained from an individual suffering from cancerous lesions within the glottal area – EGG signal (top left) and its spectral representation (bottom left), and audio signal (top right) and its spectral representation (bottom right).

2.4.2 Other Growths

Within the pathological group of "other growths of the larynx" we included all benign growths that do not directly impact the vocal folds or glottis but may still arise within the

laryngeal structures. These conditions include cases majorly unrelated to the mechanisms of human phonation, with examples including benign laryngeal cysts, and granulomas. While these growths are non-cancerous, their presence can sometimes lead to airway obstruction, difficulty swallowing, or mild voice disturbances, particularly if they exert pressure on surrounding tissues (Soni *et al.*, 2016).

The causes of such benign growths are varied and include chronic irritation, infections, or prolonged mechanical trauma (Courey *et al.*, 1996). For instance, granulomas often develop due to persistent inflammation caused by gastroesophageal reflux disease (GERD) or prolonged intubation during medical procedures. Treatment for these conditions often focuses on addressing the underlying cause and alleviating symptoms. In most cases, these benign conditions pose minimal risk to overall health but may require monitoring to prevent recurrence or complications.

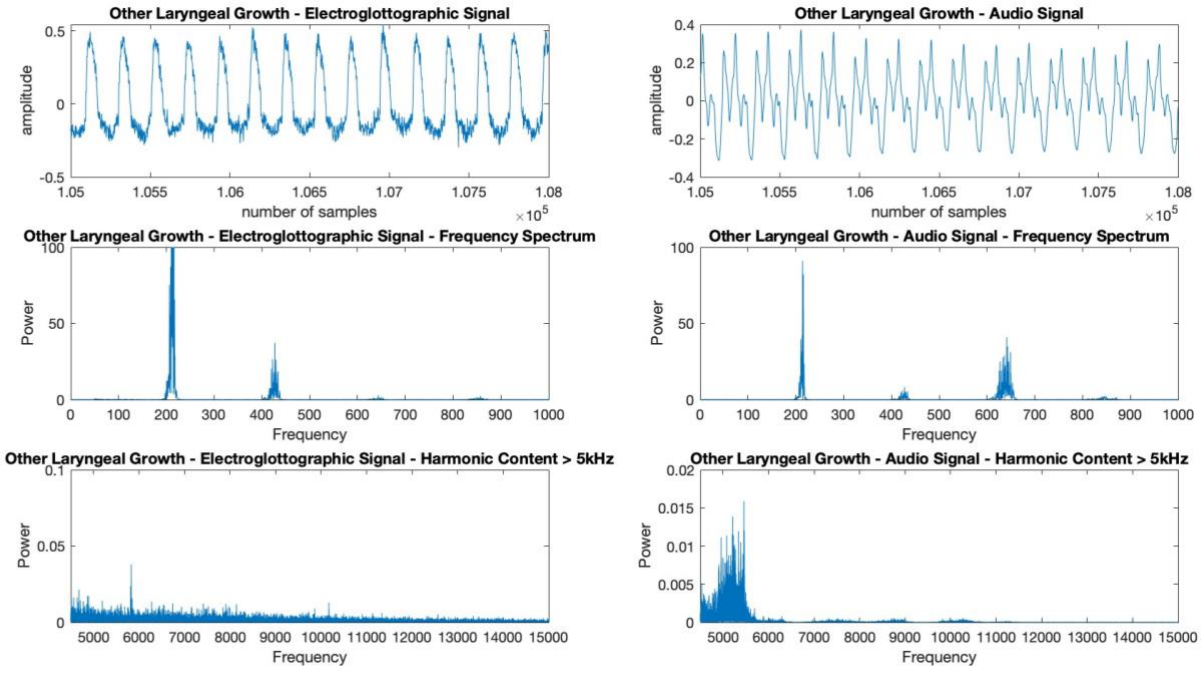

Figure 2.9: Figure 2.9: "Other Laryngeal Growths Signal". Representation of a signal obtained from an individual suffering from benign laryngeal polyps not affecting vocal folds – EGG signal (top left) and its spectral representation (bottom left), and audio signal (top right) and its spectral representation (bottom right).

Several studies indicate that benign lesions located outside of the vocal folds and the glottis area show minimal impact on the vibratory behaviour of the vocal folds (Lee *et al.*, 2019; Shrivas *et al.*, 2022). Furthermore, the benign laryngeal growths often lead to chronic inflammation, causing the co-occurrence with chronic laryngitis (Soni *et al.*, 2016; Jetté, 2016). For those reasons, as well as due to large variability of the benign growth types in the collected data, the subset of “other growths” was omitted in the final dataset setup chosen for the development of the envisaged multi-class laryngeal pathology classification system. Nevertheless, the group of other growths was still included in the development of the binary laryngeal pathology detection system.

2.4.3 Neuromuscular Disorders

Most of the neuromuscular cases investigated in this research are a direct result of an underlying damage or extreme irritation caused to the recurrent laryngeal nerve – those cases predominantly result in a permanent or temporary paralysis of the affect nerve and, consequently, vocal fold paralysis on the side of the affected nerve.

Vocal fold paralysis is a laryngeal disorder characterised by the neurologically conditioned inability of one or both vocal folds to move properly. It is most commonly caused by laryngeal nerve paralysis following an extensive infection (for instance, Lyme disease), a damage to laryngeal muscles, tumours within the lung, neck or thyroid area, or trauma induced by a surgery – most commonly, a thyroid gland surgery (Rubin and Sataloff, 2007). Due to a permanent glottic dilation caused by vocal fold paralysis, the affected person’s ability to produce voice is significantly impaired, leading to their voice sounding unnaturally breathy. Depending on the muscle tone of the paralysed vocal fold, it may constitute an airway obstruction. These cases can be treated by permanent glottic dilation via laterofixation or cordectomy (Moustafa *et al.*, 1992).

Patients with unilateral vocal fold paralysis (where only one side of the focal folds is affected) typically experience hoarseness, breathiness, and difficulty in controlling pitch and loudness during speech. When both vocal folds are paralysed, it can result in severe airway obstruction, causing breathing difficulties, especially during inhalation. This may lead to shortness of breath and respiratory distress, especially in stressful situations. Additionally, the lack of coordination between the vocal folds can lead to aspiration, where food or liquid enters the airway, posing a risk of choking and respiratory infections (Rubin and Sataloff, 2007).

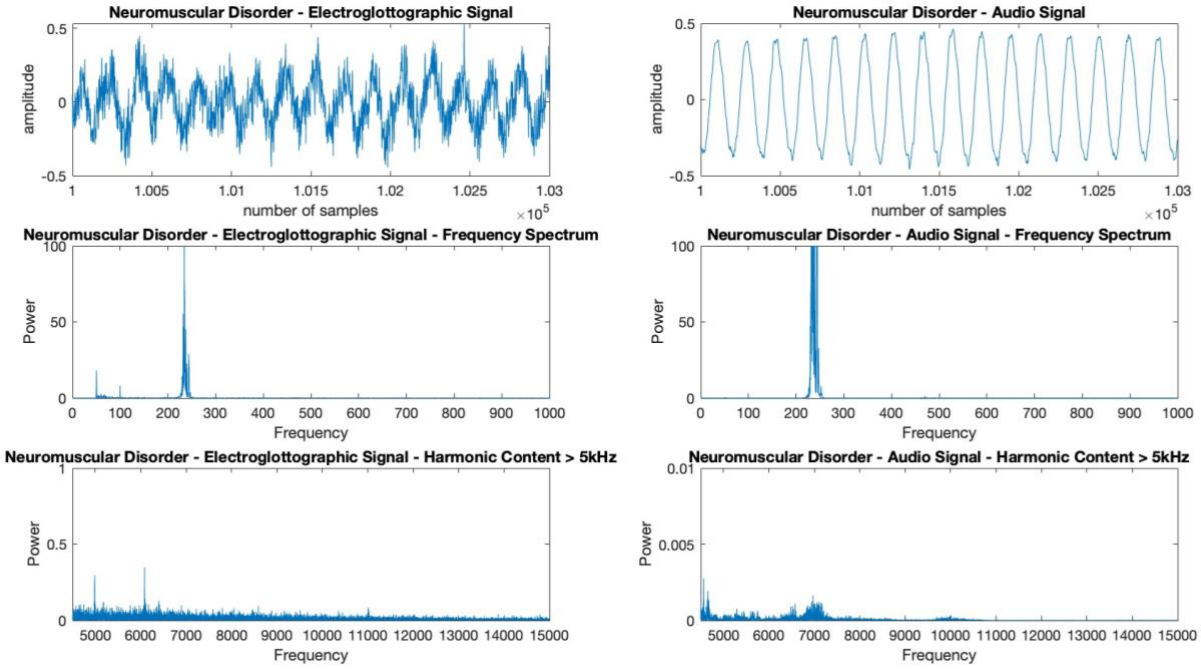

Figure 2.10: "Neuromuscular Disorder Signal". Representation of a signal obtained from an individual suffering from vocal fold paralysis – EGG signal (top left) and its spectral representation (bottom left), and audio signal (top right) and its spectral representation (bottom right).

2.4.4 Laryngitis

Laryngitis is an inflammatory condition affecting the tissues within the larynx. It is characterised by excessive swelling, redness, and irritation of the vocal cords and the surrounding tissues. This condition often leads to the alteration or loss of voice quality

(dysphonia for the alternation and aphonia in case of loss of voice – Roy, 2003), with symptoms including hoarseness, sore throat, and a persistent cough. Additionally, difficulty swallowing and the sensation of a lump in the throat are common symptoms.

Acute laryngitis is usually caused by bacterial or viral infections, such as the common cold or influenza, excessive voice use, or irritants like smoking. On the other hand, chronic laryngitis may result from long-term exposure to irritants, reflux disease, or prolonged vocal strain (Dworkin, 2008). The prolonged inflammation of the larynx, especially in the vocal cords area, often leads to vocal cord oedema, causing hoarseness and a lowered pitch of the voice (Soni *et al.*, 2016).

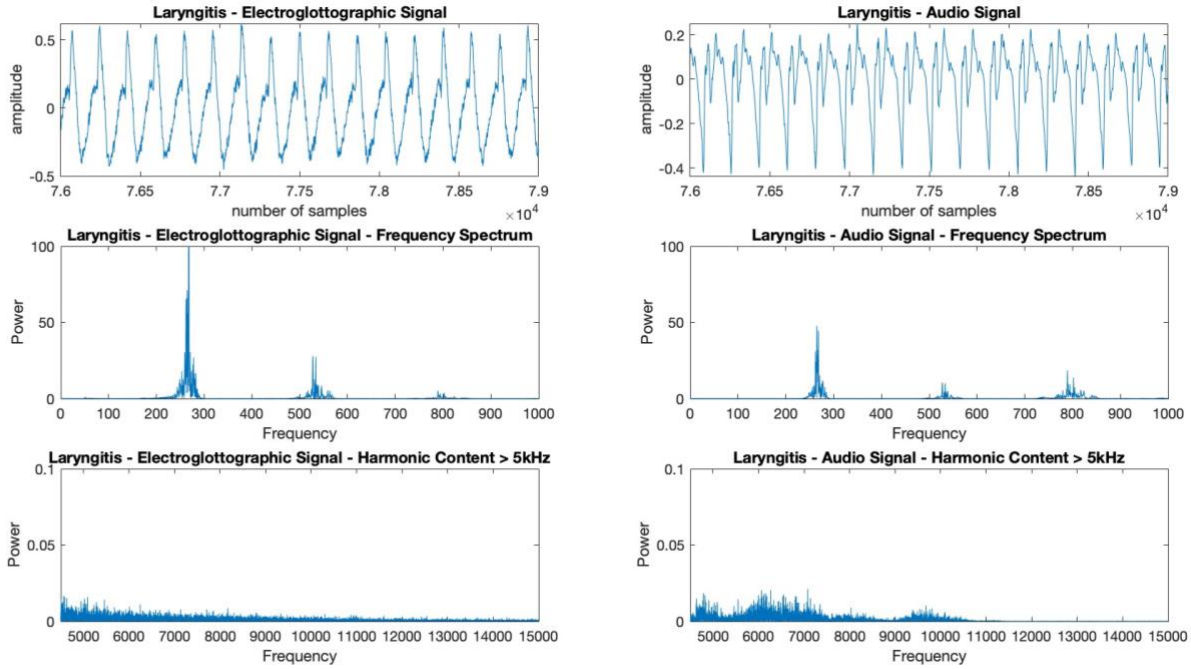

Figure 2.11: "Laryngitis Signal". Representation of a signal obtained from an individual suffering from laryngitis – EGG signal (top left) and its spectral representation (bottom left), and audio signal (top right) and its spectral representation (bottom right).

Nevertheless, the laryngitis can also co-occur with other laryngeal disorders – chronic laryngeal inflammation is commonly observed in patients suffering with malignant or benign growths of the larynx, for example Reinke's Oedema (Soni *et al.*, 2016; Jetté, 2016). Since

it often arises along with another underlying laryngeal condition, and it is present in majority of participants with benign and malignant growths recruited for the purposes of data collection in this research, it is severely difficult to classify it as a standalone condition. For that reason, the final multi-class classification model proposed in this research does not consider the laryngitis. However, as a key symptom of a pathology, laryngitis was taken into consideration during the creation of the final binary laryngeal pathology detection system.

2.4.5 Reinke's Oedema

Vocal cord oedema stands for an inflammatory state of the folds characterised by the chronic accumulation of fluid in their superficial layer (Dewan *et al.*, 2022). The state of chronic inflammation in Reinke's Oedema is caused by a prolonged irritation such as chronic smoke exposure, laryngopharyngeal reflux or intense vocal overuse, hence it is often associated with cigarette abuse. This pathology is characterised with fluid-filled masses arising within Reinke's space on both sides of the glottis. It predominantly affects adults, and it is more commonly diagnosed in women due to more noticeable lowering of the vocal pitch associated with Reinke's Oedema (Dewan *et al.*, 2022).

The abnormal thickening of the vocal folds that characterises Reinke's Oedema results in their increased mass and stiffness, which leads to a fundamental alteration in their vibratory characteristics. This, in turn, affects the quality and pitch of the voice, causing a hoarse and low-pitched vocal quality. Additionally, individuals with Reinke's Oedema may experience discomfort or pain in the throat, a sensation of fullness, and difficulties in producing clear speech (Dewan *et al.*, 2022).

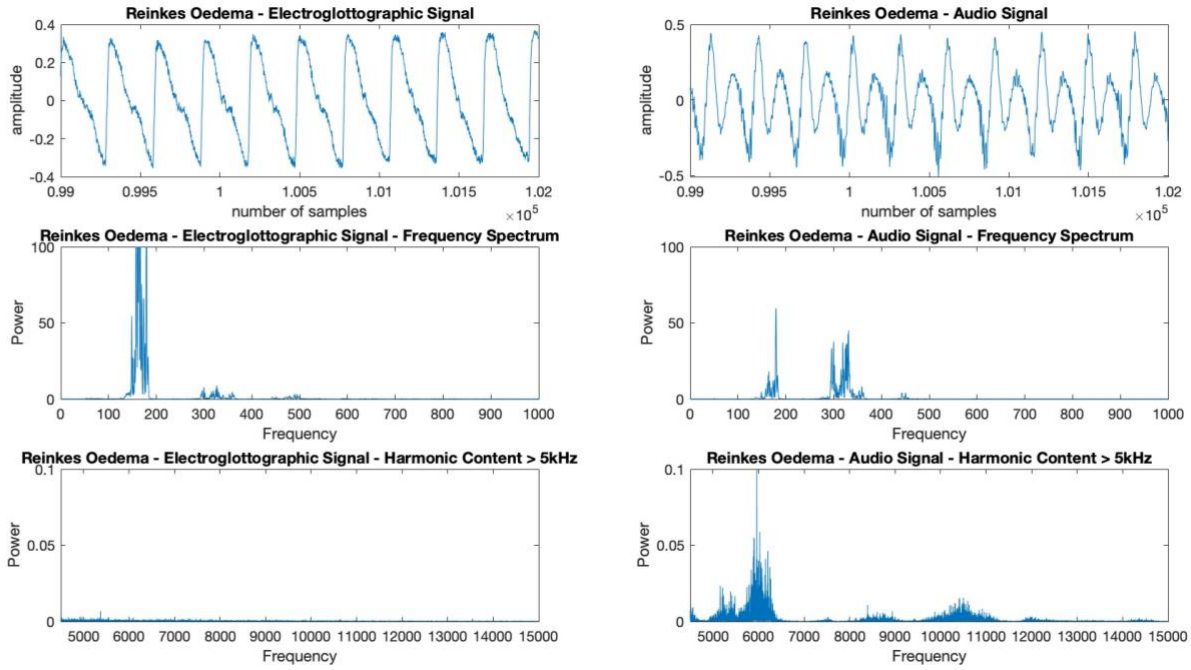

Figure 2.12: "Reinke's Oedema Signal". Representation of a signal obtained from an individual suffering from Reinke's Oedema – EGG signal (top left) and its spectral representation (bottom left), and audio signal (top right) and its spectral representation (bottom right).

In practice, the voice changes from Reinke's Oedema (e.g. chronic hoarseness and lowered pitch) often overlap with other benign vocal fold lesions or even normal voice variations. For instance, an otolaryngology review notes that vocal fold polyps, nodules, Reinke's oedema, and similar Reinke's-space lesions all "share a sub-epithelial edema" and can be difficult to distinguish from one another, even under clinical examination or histology (Bohlender, 2013). For these reasons, the Reinke's oedema subset of pathologies was not included in the final multi-class laryngeal pathology classification system. However, given its prevalence as a common laryngeal disorder, it was considered in the development of the final binary laryngeal pathology detection system.

2.4.6 Dysphonia

Functional dysphonia is one of the most prevalent symptoms of laryngeal pathologies and it is therefore referred to throughout the study. According to Martins *et al.* (2015),

functional dysphonia constitutes to over 20% of all voice-related diagnoses among adults (between 19 and 60 years of age). Nevertheless, it is not considered a condition in itself (Przysiezny and Przysiezny, 2015), instead, it constitutes a manifestation of an underlying pathology related to voice production (Roy, 2003). There are several cases of subjects diagnosed with functional dysphonia that can be found in the dataset recorded for the purposes of this study.

Functional dysphonia is characterised by vocal changes that are not related to structural or neurological abnormalities; instead, they are attributed to the improper use of the vocal apparatus. Consequently, during speech or singing, the vocal folds do not adduct appropriately, causing an irregular airflow pattern, resulting in voice disturbances or a breathy sounding voice. An instance of such a case can be glottic insufficiency – the incomplete closure of vocal folds, the risk of which increases with age (Gregory *et al.*, 2012). This condition predominantly affects the quality and function of the voice without an underlying organic pathology. It is often associated with a range of other vocal symptoms such as hoarseness, breathiness, pitch breaks, and vocal fatigue.

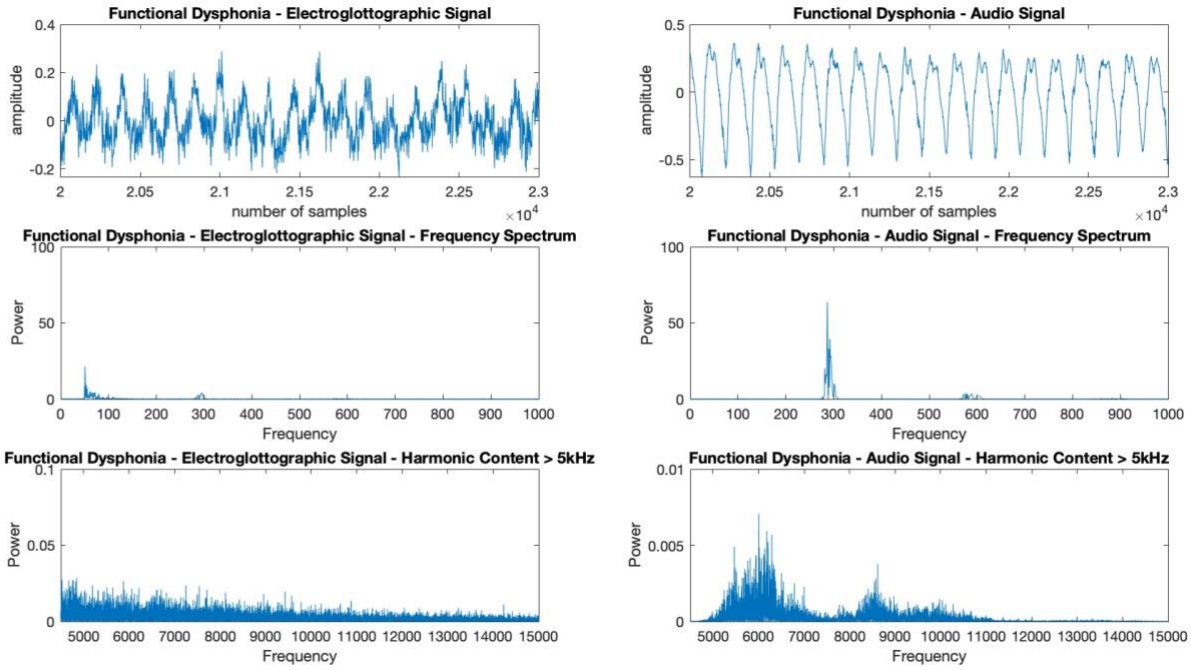

Figure 2.13: "Dysphonic Signal". Representation of a signal obtained from an individual suffering from dysphonia – EGG signal (top left) and its spectral representation (bottom left), and audio signal (top right) and its spectral representation (bottom right).

Since dysphonia is considered a symptom rather than a standalone diagnosis (Przysiezny and Przysiezny, 2015), this subset of pathologies was not considered during the development of the final multi-class laryngeal pathology classification system. However, as a significant symptom of an underlying laryngeal pathology, it was considered during the development of the final binary laryngeal pathology detection system.

Chapter 3

Review of Literature on Laryngeal Pathology Classification Systems

Historically, audio-based assessment methods have been instrumental in vocal tract pathology detection, leveraging the distinct acoustic characteristics of pathological voices. The recent expansion of novel digital signal processing and machine learning approaches opened new avenues for the development of sophisticated deep learning models capable of distinguishing between pathological and healthy signals with high accuracy. Despite notable advancements, existing classification systems often struggle with multi-class discrimination, particularly when differentiating between conditions with overlapping acoustic features.

Similarly, electroglottography has emerged as a valuable non-invasive tool for assessing vocal fold function through bioimpedance measurements. Initially developed for phonatory function analysis, EGG has recently been readopted in research and clinical settings. Nevertheless, challenges persist when using laryngeal bioimpedance as the sole modality is laryngeal pathology classifiers. Studies have shown that EGG's accuracy (for exact definition of accuracy, see section *8.1. Methods of Results Assessment*) in identifying laryngeal disorders in a unimodal approach often falls below 60%, which suggests performance comparable to a random assignment (Borsky *et al.*, 2017; Miliaresi *et al.*, 2022; Islam *et al.*, 2022).

Recent years have seen growing interest in deep learning models for laryngeal pathology detection, including those that bypass handcrafted feature extraction or statistical evaluation methods by learning representations directly from raw data. These methods include autoencoders – a class of neural networks designed to compress input into a meaningful

representation, learn compressed features, and decode it attempting to replicate input's original state in an unsupervised manner. Widely used in image and video analysis (Markatopoulou et al., 2018), autoencoders have also been explored in medical signal and image analysis; Praveen et al. (2018) demonstrated the use of stacked sparse autoencoders for lesion segmentation in brain MRI, while Al Rahhal et al. (2016) employed a similar architecture to extract discriminative features for electrocardiogram (ECG) classification. In cancer diagnostics, Adem et al. (2019) used stacked autoencoders with a Softmax layer for cervical cancer detection. Nevertheless, the application of autoencoders in laryngeal signals remains limited, particularly for cancerous and precancerous lesion detection.

A deep learning method capable of mapping raw signals directly to output labels without intermediate feature engineering are neural networks with end-to-end architecture (Bounareli et al., 2025). This approach, commonly employed in environmental sound classification, achieved state-of-the-art performance of 89% accuracy using a 1D convolutional network (Abdoli et al., 2019). Similarly, end-to-end pipelines have been successfully applied to pathological voice data (e.g., Liu et al., 2023). Despite these advances, their role in laryngeal pathology detection remains underexplored, with most existing works focusing on sustained vowel phonation or laryngeal imaging rather than continuous speech or bioimpedance signals. This gap highlights both the potential and the challenges of adopting end-to-end or self-supervised frameworks for medical voice and laryngeal bio-signal analysis.

The following chapter provides a comprehensive review of the state-of-the-art methodologies for laryngeal pathology classification, focusing on the use of audio signals and laryngeal bioimpedance measurements as diagnostic measures. It examines the evolution of these modalities in the assessment of human phonation, encompassing both statistical and deep learning approaches. By highlighting the strengths and limitations of

audio and laryngeal bioimpedance as singular modalities, the gaps are identified in current methodologies, providing justification for the investigation of multimodality in laryngeal pathology classification. Critically examining these various aspects, this chapter establishes the foundation for the novel approaches presented in subsequent sections of this work.

3. REVIEW OF LITERATURE ON LARYNGEAL PATHOLOGY CLASSIFICATION SYSTEMS

3.1. AUDIO IN DIAGNOSTICS OF LARYNGEAL PATHOLOGIES

The use of sound in diagnostics dates to the early 19th century, with René Laennec's invention of the stethoscope, revolutionising auscultation practices (Bishop, 1980). Since then, sound has become an essential tool in diagnostics, not only in cardiology and pulmonology but also in the assessment of vocal tract disorders. Lung auscultation, for instance, remains a crucial part of respiratory system diagnostic procedures, where sounds such as rales, rhonchi, stridor, and wheezing, are associated with specific lung and bronchi conditions (Wilkins *et al.*, 1990). Various digital diagnostic systems have since been developed to analyse respiratory sounds, some achieving remarkable results (Aykanat *et al.*, 2017; Grzywalski *et al.*, 2019; Alam *et al.*, 2022).

Similarly, vocal tract pathologies, including dysphonia, nodules, and polyps, exhibit distinct acoustic features (Peng *et al.*, 2007; Henríquez *et al.*, 2009). These features have been harnessed in voice-based diagnostic digital systems to enable non-invasive monitoring and assessment (Harar *et al.*, 2017; Kuo *et al.*, 2023). However, many existing systems focus on binary classification – distinguishing healthy voices from pathological ones (Mohammed *et al.*, 2020) – or addressing broader pathology groups rather than specific conditions. While this approach simplifies classification, it limits the ability to diagnose pathologies with similar audio features – for instance, vocal fold paralysis and dysphonia, or laryngeal benign cyst and a malignant glottal lesion.

The following section provides an overview of existing laryngeal pathology classification systems, with the particular focus on unimodal systems utilising audio as the data modality. The review is structured into two parts: statistical approaches (including non-deep-learning

machine learning methods like Support Vector Machines (SVM)) and deep learning approaches.

3.1.1 Statistical and Machine Learning Approaches

Statistical methods have been pivotal in analysing acoustic features to distinguish between healthy and disordered voices. One significant early work investigated telephone-based classification of voice pathologies into four categories: normal, neuromuscular pathology, physical pathology, and mixed pathology. Using pitch perturbation, amplitude perturbation, and harmonic-to-noise ratio, Linear Discriminant Analysis (LDA) achieved 96.1% accuracy for normal samples and 92.5% for pathological samples. Upon 10-fold cross-validation implemented in this study, the results indicated 87% detection accuracy for neuromuscular conditions, 78% for physical conditions, and 61% for mixed pathologies (Moran *et al.*, 2006).

Another very popular method applied for the classification of pathological voices are the Support Vector Machines (SVMs). Peng *et al.* (2007) applied Principal Component Analysis (PCA) to extract features from sustained vowel /a/ and classified them using SVM, reaching 98.1% accuracy. The dataset used in this research was the Massachusetts Eye and Ear Infirmary (MEEI), released in 1994 by the MEEI Voice and Speech Lab in collaboration with KayPENTAX Corp (Massachusetts Eye and Ear Infirmary, 1994). The accuracy reached 97% at the implementation of 17 principal components (Peng *et al.*, 2007). Similarly, Arias-Londoño *et al.* (2010) combined SVM with Gaussian mixture models, achieving 98.23% accuracy for binary classification of healthy versus pathological signals based on sustained phonation of a vowel /a/. Two different feature sets were implemented in the study; one comprising statistical and frequency domain including parameters such as harmonic-to-noise ratio (HNR) and MFCCs, another set derived from nonlinear analysis of embedding

attractors. Lastly, Markaki and Stylianou (2011) explored modulation spectral features derived from a sustained vowel /a/ from the MEEI database. According to the results, the modulation spectral features were superior to traditional MFCCs for most pathology discrimination tasks, except for paralysis and non-paralysis distinguishing. Overall, the results yielded the classification accuracy of 94.1% with a confidence interval of ±0.28%.

Al-Nasheri *et al.* (2017) used Multidimensional Voice Program (MDVP) parameters for binary classification across several databases, achieving accuracies up to 99.8% and 99.25% for detection and classification, respectively. The MDVP include acoustic features such as fundamental frequency ($f0$), jitter, shimmer, harmonic-to-noise ratio (HNR), voice turbulence index, as well as amplitude tremor intensity index. The investigated pathologies were vocal fold cysts, vocal fold paralysis, and vocal fold polyps. For the classification between pathological categories, a 10-fold, cross-validation approach was utilised in the first study. The second study covered different frequency bands to assess their impact on the detection and classification processes. For that, a correlation function was applied. According to the study, the most conducive frequency bands for both detection and classification of laryngeal pathologies fall within the range of 1000 to 8000 Hz. Although high accuracy was reported in case of both research pieces, the studies relied on binary comparisons rather than true multi-class classification, which remains an underexplored area in statistical approaches.

Overall, in laryngeal pathology classification, the statistical and machine learning methods favour binary classification. Furthermore, there is an evident lack of research on continuous speech signals rather than sustained phonation, since large majority of statistical methods rely on recordings of sustained vowel /a/ obtained from the MEEI database (Moran *et al.*, 2006; Peng *et al.*, 2007; Arias-Londoño *et al.*, 2010; Markaki and Stylianou, 2011; Al-Nasheri *et al.*, 2017).

3.1.2 Deep Learning Approach

One of the first laryngeal pathology detection systems, fully relying on the implementation of audio signals and deep learning, was the model proposed by Godino-Llorente and Gómez-Vilda (2004), utilising Learning Vector Quantisation and the Multi-layer Perceptron. The system relied on the application of MFCCs derived from audio recordings of sustained vowel /a/, obtained from the MEEI dataset. While only capable of binary classification between pathological and healthy signals, the Learning Vector Quantisation yielded high results of 96% accuracy (Godino-Llorente and Gómez-Vilda, 2004). Similarly, Henríquez *et al.* (2009) used a Feedforward Neural Network with features including first and second order Rényi entropies, the correlation entropy, and Shannon entropy (Henríquez *et al.*, 2009). The results yielded 82.47% for the custom database, and 99.69% for the MEEI database.

The investigation of sequential characteristics of pathological versus healthy audio was pursued by Harar *et al.* (2017) with the application of a deep learning model that combines CNN with recurrent layers of Long-Short-Term-Memory network (LSTMs). Fed with raw audio recordings of sustained vowel /a/ from the Saarbruecken Voice Database (SVD), the proposed model achieved 71.36% accuracy, 65.04% sensitivity, and 77.67% specificity on 206 validation files, and 68.08% accuracy, 66.75% sensitivity, and 77.89% specificity on 874 testing files. Wang *et al.* (2022) advanced this approach to multi-class arrangement, classifying functional dysphonia, neoplasm, phonotrauma, and vocal palsy. The best performing model proposed by Wang *et al.* used Bi-Directional LSTMs (BiLSTMs) in combination with MFCCs derived from both continuous speech and sustained phonation. The model reached 89.27% accuracy. Notably, their study demonstrated the superior performance of speech over sustained vowels for feature extraction.

Recent research of laryngeal pathology classification with implementation of deep learning proposed a model combining CNN with a domain adversarial training (DAT) module

(Kuo *et al.*, 2023). The DAT is applied to enhance the model's ability to classify in noisy environments (Kuo *et al.*, 2023). Three categories of signals were investigated: heathy, affected by neoplasm, and affected by benign structural diseases. The system uses the factorised CNNs as a feature extractor, followed by the CNN-DAT combination as the classifier. The results yielded the accuracy of 88% for healthy signals detection, 79% for neoplasm, and 72% for structural conditions.

Chen and Chen (2022) explored the use of sparse stacked autoencoders fed with 12 MFCCs for the detection of pathological voices, achieving 98.6% accuracy in binary classification. More recently, Liu et al. (2023) developed an end-to-end deep learning system trained on the Saarbruecken Voice Database, achieving promising results for distinguishing laryngitis from hyperfunctional dysphonia and healthy voices – F1 score for laryngitis and healthy varied between 0.77 and 0.87, however, with significantly lower F1 score for hyperfunctional dysphonia – 0.49-0.56. While these works demonstrate the feasibility of the methods, they are constrained by reliance solely on sustained vowels and by databases such as Saarbruecken Voice Database that contains inconsistencies (for details, see section *4.4.1 Limitations of SVD*). Transfer learning from large audio resources such as Audio Set (Gemmeke et al., 2017) has also been widely explored, enabling models pre-trained on environmental and speech data to be fine-tuned for more specialised tasks. Yao et al. (2023) investigated the use a convolutional autoencoder and transfer learning for classification of essential tremor of voice and abductor and adductor spasmodic dysphonia, and healthy participants. The study achieved the maximum accuracy of 85.3% in differentiating healthy, tremor, and one type of dysphonia, but only 65.3% in classification of all investigated pathologies at once.

Furthermore, the exploration of alternative feature extraction techniques, such as the ERB spectrum-derived features (Zhou *et al.*, 2022; Kumar *et al.*, 2023; Islam *et al.*, 2022),

has yielded promising outcomes. Zhou *et al.* (2022) proposed a system employing Gammatone Spectral Latitude for feature extraction, combined with classification methods including Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Random Forest (RF). This system achieved a discrimination accuracy of 99.6% for distinguishing between specific pathology variants. However, cross-database testing revealed challenges in generalisability, with classification accuracy dropping to approximately 70% and further decreasing to 55.7% when tested on the Saarbruecken Voice Database. The average accuracy for detecting “structural” conditions, encompassing certain precancerous and cancerous categories within SVD, was calculated at 59.61%. Notably, the study focused exclusively on sustained phonation recordings, without investigating the use of continuous speech signals.

Overall, the deep learning approach to vocal tract pathology classification yields good results and remains highly promising for future investigation. Nevertheless, there is very little research within classification of particular laryngeal diseases, with the emphasis on malignant growths and cancerous and precancerous lesions. Integrating audio data with other modalities, such as electroglottographic measurements, offers a potential path forward (Tomaszewska and Georgakis, 2023).

3.1.3 Summary of Audio-based Methods

Research on audio-based voice pathology classification has primarily relied on statistical and machine learning approaches to process features such as pitch perturbation, amplitude perturbation, harmonic-to-noise ratio, and Mel spectrum-derived features (Godino-Llorente and Gomez-Vilda, 2004; Moran *et al.*, 2006; Peng *et al.*, 2007; Arias-Londoño *et al.*, 2010; Markaki and Stylianou, 2011; Hemmerling *et al.*, 2016; Al-Nasheri *et al.*, 2017; Wang *et al.*, 2023). Statistical methods, including Linear Discriminant Analysis and Support Vector

Machines (SVM), have achieved high accuracies, such as 98.1% (Peng *et al.*, 2007) and 98.23% (Arias-Londoño *et al.*, 2010), particularly in binary classification. Feature extraction techniques like Principal Component Analysis and Gammatone Spectral Latitude have also shown promise, though generalisability across databases remains a challenge (Hemmerling *et al.*, 2016; Zhou *et al.*, 2022).

Recent advances in deep learning have enabled classification of laryngeal pathologies based on structural and neuromuscular distinctions. Models such as Bi-Directional Long-Short Term Memory Networks (BiLSTM) and Convolutional Neural Networks (CNN) have reported accuracies of 89.27% and up to 88%, respectively (Wang *et al.*, 2022; Kuo *et al.*, 2023). While autoencoders, end-to-end pipelines, and transfer learning approaches have demonstrated promise in voice analysis, most existing studies rely on sustained phonation and datasets of limited reliability. Furthermore, true multi-class classification remains underexplored, with most systems relying on binary comparisons or achieving limited accuracy in multi-class scenarios (Markaki and Stylianou, 2011; Al-Nasheri *et al.*, 2017; Borsky *et al.*, 2017; Islam *et al.*, 2022; Liu et al., 2023; Yao et al., 2023).

Most research to date has focused on sustained phonation due to its stability and reduced articulatory variability (Rosa *et al.*, 1999; Henríquez *et al.*, 2009; Markaki and Stylianou, 2011; Islam *et al.*, 2022; Zhou *et al.*, 2022; Kuo *et al.*, 2023), achieving high accuracy but limiting the detection of patterns revealed in dynamic, continuous speech. The integration of audio with modalities like electroglottography offers potential for more robust systems (Tomaszewska and Georgakis, 2023). Addressing gaps in the use of continuous speech and exploring novel feature extraction methods remain crucial for advancing the field.

Table 3.1 provides a summary of selected previous work on audio-based classification of laryngeal pathologies.

Table 3.1 Summary of seminal published work on audio-based laryngeal pathology classification.

REFERENCE	DATASET	OBJECTIVE	METHODS	FINDINGS
Moran *et al.*, 2006.	631 speakers (573 pathological, 58 control) 151 pathological speakers for pairwise classification between control and specific pathology (56 neuromuscular, 56 "mixed", 39 "physical"). Male and Female.	Binary classification between control and pathological (neuromuscular, "physical", "mixed"). Testing two types of audio – recordings collected in a controlled environment, as well as the telephone-quality recordings.	DATA: Massachusetts Eye and Ear Infirmary (MEEI) – Disordered Voice Database Model 4337 – sustained phonation vowels /a/. FEATURES: Pitch perturbation, amplitude perturbation and harmonic-to-noise ratio (HNR). CLASSIFICATION: Linear discriminant analysis.	ACCURACY: Controlled environment recordings: 89.10%. Telephone-quality recordings: 74.2% Control vs neuromuscular: 87.27%, Control vs "physical": 77.97%, Control vs "mixed": 61.08%
Peng *et al.*, 2007.	216 speakers (177 pathological, 39 control). Male and Female.	Binary classification between control and pathological.	DATA: Massachusetts Eye and Ear Infirmary (MEEI) – sustained phonation vowels /a/. FEATURES: Acoustic features (Multidimensional Voice Program – MDVP), and Principal Component Analysis (PCA). CLASSIFICATION: Support Vector Machine (SVM).	ACCURACY: 98.10%
Henríquez *et al.*, 2009.	Multiquality Database – 142 speakers (57 pathological, 85 control), MEEI Database – 226 speakers (173 pathological, 53 control) Male and Female.	Binary classification between control and pathological.	DATA: Multiquality Database – various sustained phonation vowels, Massachusetts Eye and Ear Infirmary (MEEI) – sustained phonation vowels /a/. FEATURES: Quantification of audio recordings through: first and second order Rényi entropies, the correlation entropy, the correlation dimension, the value of the first minimum of mutual information function, Shannon entropy. CLASSIFICATION: Multilayered Feedforward Neural Network.	ACCURACY: Multiquality database: 82.47%, MEEI: 99.69%.
Arias-Londoño *et al.*, 2010.	226 speakers (173 pathological, 53 control) Male and Female.	Binary classification between control and pathological.	DATA: Massachusetts Eye and Ear Infirmary (MEEI) – sustained phonation vowels /a/. FEATURES: Harmonics-to-noise ratio (HNR), normalised noise energy (NNE), glottal to noise excitation ratio (GNE), as well as 12 MFCCs. CLASSIFIER: Fusion of Gaussian mixture models (GMM) and Support Vector Machine (SVM).	ACCURACY: 98.23%
Markaki and Stylianou, 2011.	226 speakers (173 pathological – 20 "nodules", 20 "polyps", 26 "keratosis", 22	Binary classification between control and pathological, and binary pairwise discrimination between	DATA: Massachusetts Eye and Ear Infirmary (MEEI) – sustained phonation vowels /a/.	Modulation Spectral Features outperforming the MFCCs.

	"adductor", 71 "paralysis", remaining 14 unspecified, and 53 control) Male and Female.	individual pathologies ("nodules", "polyps", "keratosis", "adductor"), as well as binary pairwise discrimination between "nodules", "polyps", "keratosis", and "adductor" collectively against "paralysis".	FEATURES: Modulation Spectral Features, compared to MFCCs. CLASSIFIER: Support Vector Machine (SVM).	Results for Modulation Spectral Features: ACCURACY: Control vs pathological: 94.1%. AREA UNDER THE ROC CURVE: "Polyp" vs "adductor": 0.9585 "Polyp" vs "keratosis": 0.9359 "Polyp" vs "nodules": 0.9428 "Adductor" vs "nodules": 0.9578 "Adductor" vs "keratosis": 0.9949 "Keratosis" vs "nodules":0.9527 "Paralysis" vs others: 0.7648
Hemmerling, *et al.*, 2016.	900 speakers (450 pathological, 450 control) Male and Female.	Binary classification between control and pathological.	DATA: Saarbruecken Voice Database (SVD) – various sustained phonation vowels. FEATURES: Acoustic features, and Principal Component Analysis (PCA). CLASSIFIER: Random Forest Classifier.	ACCURACY: 99% accuracy.
Al-Nasheri *et al.*, 2017.	AVPD Database – 193 speakers (75 pathological – 13 "cysts", 32 "paralysis", 30 "polyps", and 118 control), MEEI Database – 148 speakers (95 pathological – 10 "cysts", 66 "paralysis", 19 "polyps", and 53 control), SVD Database – 506 speakers (6 "cysts", 195 "paralysis", 43 "polyps", and 262 control). Male and Female.	Binary pairwise discrimination between pathologies ("cysts", "paralysis", "polyps").	DATA: Massachusetts Eye and Ear Infirmary (MEEI) – sustained phonation vowels /a/. Saarbruecken Voice Database (SVD) – various sustained phonation vowels. Arabic Voice Pathology Database (AVPD) – various sustained phonation vowels. FEATURES: Acoustic features (Multidimensional Voice Program – MDVP). CLASSIFIER: Fisher discrimination ratio.	ACCURACY: Cysts vs other: MEEI: 88.89%, SVD: 97.5%, AVPD: 82.86%. Paralysis vs other: MEEI: 65.56%, SVD: 79.17%,. AVPD: 57.14%. Polyps vs other: MEEI: 30%, SVD: 82.08%, AVPD: 60%..
Harar *et al.*, 2017.	1166 speakers (583 pathological, 583 control). Male and Female.	Binary classification between control and pathological.	DATA: Saarbruecken Voice Database (SVD) – sustained phonation vowels /a/. FEATURES: Audio files fed into the network in 64 ms long segments with 30 ms overlap. CLASSIFIER: Convolutional Neural Networks with Long-Short Term Memory Networks.	ACCURACY: 71.36 %
Wang *et al.*, 2022.	1045 speakers (all pathological – 100 "functional dysphonia", 103 "neoplasm", 718 "phonotrauma", 124 "vocal palsy"). Male and Female.	Multi-class discrimination between 4 classes of common aetiology pathologies ("functional dysphonia", "neoplasm", "phonotrauma", and "vocal palsy"). No control class of healthy participants.	DATA: Far Eastern Memorial Hospital (FEMH) database – continuous speech recordings. FEATURES: Mel-Frequency Cepstral Coefficients (MFCCs) CLASSIFICATION:	ACCURACY: 89.27% maximum.

			Bi-Directional Long-Short Term Memory Network (BiLSTM).	
Zhou *et al.*, 2022.	MEEI Database – 265 speakers (212 pathological – 92 neuromuscular, 120 “structural”, and 53 control), SVD Database – 1181 speakers (494 pathological – 287 neuromuscular, 207 “structural”, and 687 control), HUPA Database – 398 speakers (212 pathological – 31 neuromuscular, 128 “structural”, and 239 control). Male and Female.	Multi-class discrimination between 2 classes of common aetiology pathologies plus control (neuromuscular, “structural”, and control).	DATA: Massachusetts Eye and Ear Infirmary (MEEI), Saarbruecken Voice Database (SVD), Hospital Universitario Prıncipe de Asturias (HUPA). FEATURES: Gammatone Spectral Latitude Features (GTSLs) compared to GTCCs. CLASSIFICATION: Multilayer Perception (MLP), Support Vector Machine (SVM) and Random Forest (RF).	GTSLs outperforming the GTCCs. Results for GTSLs: ACCURACY: Control vs pathological: 99-100%. Neuromuscular vs “structural” vs control: MEEI: 99.6%. SVD: 89.9%, HUPA: 97.4%.
Chen and Chen, 2022.	208 speakers (151 pathological, and 57 control). Male and Female.	Binary classification between control and pathological.	DATA: Voice Icar fECerico II Database of PhysioNet – 5-second long sustained phonation vowels. FEATURES: Mel-Frequency Cepstral Coefficients (MFCCs) CLASSIFICATION: 2-layer stacked sparse autoencoder with softmax layer.	ACCURACY: 98.6%
Kuo *et al.*, 2023.	523 speakers (415 pathological – 112 “neoplasms”, 303 “benign structural diseases”, and 108 control). Male and Female.	Multi-class discrimination between 2 classes of common aetiology pathologies plus control (“neoplasm”, “benign structural diseases”, and control). Testing recordings gathered in two types of environments – the clinical environment and the noisy real-world environment. Additional testing completed on two-stage classification with implementation of CNN-based noise detection.	DATA: Saarbruecken Voice Database (SVD) – sustained phonation vowels /a/, /i/, /u/. CLASSIFIER: End-to-end Convolutional Neural Networks with Long-Short Term Memory Networks.	UNWEIGHTED AVERAGE RECALL (UAR): Clinical environment: 80%. Targeted noisy real-world environment: 72%. UAR OF TWO-STAGE CLASSIFIER: Clinical environment: 84%. Targeted noisy real-world environment: 71%.
Liu *et al.*, 2023.	1021 speakers (334 pathological – 207 “hyperfunctional dysphonia”, 127 “laryngitis”, and 687 control). Male and Female.	Multi-class discrimination between 3 classes (“hyperfunctional dysphonia”, “laryngitis”, and control).	DATA: Far Eastern Memorial Hospital (FEMH) database – various sustained phonation vowels. FEATURES: Convolutional Neural Networks (CNN) CLASSIFICATION: CNN with additional implementation of domain adversarial training (DAT) module.	UNWEIGHTED AVERAGE RECALL (UAR): Clinical environment: 80%. Targeted noisy real-world environment: 72%. UAR OF TWO-STAGE CLASSIFIER: Clinical environment: 84%. Targeted noisy real-world environment: 71%.
Yao *et al.*, 2023.	134 speakers (134 pathological – 44 “adductor spasmodic dysphonia”, 45 “abductor spasmodic dysphonia”, 45 “tremor”, and 145 control). Male and Female.	Multi-class discrimination between 4 classes (“adductor spasmodic dysphonia” (ADSD), “abductor spasmodic dysphonia” (ABSD), “tremor” (ETV), and control), as well as in 3 class setup.	DATA: Vanderbilt University Medical Center (VUMC) for pathological data – sustained phonation of /a/ and /i/. International Dialect of English Archives (IDEA) for	ACCURACY: Maximum between control, ABSD and ETV: 85.3% Maximum between control, ADSD and ETV: 77.7%

			control recordings – Rainbow Passage and speech. FEATURES: Mel spectrograms and time-series waveform files. CLASSIFICATION: CNN with additional implementation of domain adversarial training (DAT) module.	Maximum between all: 65.3%

3.2. BIOIMPEDANCE IN DIAGNOSTICS OF LARYNGEAL PATHOLOGIES

Electroglottography (EGG), introduced in the 1940s for pulse frequency measurement, quickly emerged as an important tool for investigation of the vocal tract function. In the 1950s, Fabre developed the “high-frequency glottograph”, pioneering the application of EGG in voice pathology research (Fabre, 1957). Further validation by van Michel *et al.* (1970) and subsequent advancements by Fourcin and Abberton in the 1970s – including the crucial development of the laryngograph (Abberton and Fourcin, 1972; Fourcin, 1974) – expanded EGG’s role in laryngeal assessment and glottal phase analysis. By the 1980s, Childers’ work on distinguishing pathological larynges highlighted both the potential, as well as the limitations of EGG, particularly regarding its representation of the glottal area (Childers *et al.*, 1986; Childers *et al.*, 1987). Colton and Conture (1990) later emphasised its accuracy in capturing fundamental frequency ($f0$), while, in his review, Baken (1992) underlined its unique contributions to phonatory function analysis. Technological improvements in spatial resolution during the 1990s further enhanced the electroglottograph’s utility in tracking laryngeal movement (Rothenberg, 1992; Rothenberg, 2002). In the early 2000s, EGG applications broadened to include speech processing and medical diagnostics, renewing interest in its use for laryngeal assessment and related pathology classification.

The full overview of the history of electroglottography is available under the *Appendices* section of this thesis (*Table 12.2: History of Electroglottography*).

Due to its non-invasive and cost-effective nature, EGG remains an attractive option for examination of laryngeal function (Jiang *et al.*, 1998; Carding *et al.*, 1999; Ritchings *et al.*, 2001). Nevertheless, it also presents several limitations; factors such as electrode placement and skin contact significantly influence electroglottographic signal quality, making its standardisation challenging. Inconsistencies in waveform interpretation and variability in measurements can complicate analysis, especially when comparing across studies. The key EGG parameters essential for understanding vocal fold dynamics include (Henrich *et al.*, 2004; Henrich *et al.*, 2005):

- Glottal Closure Instants (GCIs): The moment of sudden vocal fold closure, marking the start of the closed phase.
- Glottal Opening Instants (GOIs): The moment vocal folds begin to reopen due to air pressure and muscle tension, signalling the start of the open phase.
- Fundamental Period: the interval between consecutive GCIs, representing one cycle of vocal fold oscillation
- Open Phase: The duration between GOI and the subsequent GCI, when vocal folds are abducted.
- Closed Phase: The duration between GCI and the subsequent GOI, when vocal folds are adducted.
- Open Quotient: The ratio of the open phase to the fundamental period.
- Closed Quotient: The ratio of the closed phase to the fundamental period.
- dEGG (Differentiated EGG): The first derivative of the EGG signal, characterised by distinct positive and negative peaks linked to glottal closure and opening (Henrich *et al.*, 2004; Henrich *et al.*, 2005).

It should be recognised that the above parameters are considered theoretical and subject to variability depending on the analytical approach. Interpretations should be made with caution, particularly when conducting statistical analyses based on these measures.

3.2.1 Statistical and Machine Learning Approach

Early systems for digital EGG signal classification heavily relied on statistical methods; those included perturbation measures (Childers and Bae, 1992; Hosokawa *et al.*, 2014), quotients (Jiang *et al.*, 1998), glottal instants (Deshpande and Manikandan, 2017), and harmonic content analysis (Ritchings *et al.*, 2001). These were predominantly used alongside various statistical classifiers, including Mann-Whitney U test (Nacci *et al.*, 2020), Random Forest, and GMM – Gaussian Mixture Model (Borsky *et al.*, 2017).

With the application of perturbation analysis, Linear Predictive coding and Vector Quantisation, Childers and Bae (1992) proposed a pioneering model for digital EGG evaluation and detection of pathological signals, achieving the accuracy of 75.9% and 69% for audio recordings (of a vowel /i/ phonation sustained over 2 seconds), and EGG, respectively. The authors postulated that laryngeal disorders are more apparent in stable cycles interspersed with the unstable fragments.

Jiang *et al.* (1998) combined EGG with photoglottogram (PGG) data to detect vocal fold paralysis, achieving the respective accuracy of 43%, 73%, and 57% for the detection of healthy, recurrent paralysis, and superior paralysis, and 64% of classification accuracy between healthy and pathological. The authors highlighted the need for larger datasets and improved data acquisition methods, suggesting EGG as a sole data modality is unreliable. Contrarily, Hosokawa *et al.* (2014) demonstrated that perturbation parameters derived from EGG were more reliable than those from audio, particularly for identifying mild dysphonia,

with receiver operating characteristic analysis confirming higher accuracy for EGG in dysphonic classification.

The development of algorithms for glottal event detection further advanced EGG-based analysis. The SIGMA algorithm (Thomas and Naylor, 2009), based on wavelet transforms and Gaussian mixture modelling, achieved high accuracy for detecting glottal closure (99.47%) and opening (99.35%). Deshpande and Manikandan (2017) proposed a multi-stage system incorporating artifact removal, glottal instant detection, and parameter extraction, achieving 94.38% accuracy in noise-free environments and 95.06% in noisy conditions. While both systems excel at extracting glottal parameters, they remain untested in full laryngeal pathology classification systems.

To explore dysphonia-related features, Nacci *et al.* (2020) introduced a Variability Index (VI) derived from EGG amplitude and speed, finding that specific glottal cycle stages, such as VI-Q2 (vocal fold contact), were highly indicative of pathology. Their statistical analysis, using Kruskal-Wallis test and the Mann-Whitney U test corrected with Bonferroni, identified significant differences in VI stage of glottal cycle for healthy versus pathological voices, with a specificity of up to 77.8%.

Borsky *et al.* (2017) compared statistical classifiers like Random Forest (RF), SVM, and Gaussian Mixture Models (GMM) for voice quality classification using audio, EGG, and glottal inverse filtered waveforms. For comparison purposes, a simple deep neural network (DNN) classifier of feed-forward architecture was also investigated, achieving an accuracy lower than RF and SVM. Although not directly pertaining to the pathologies of the larynx, the study provided valuable insights into the varying laryngeal bioimpedance depending on a voice mode – more specifically breathy, strained, and rough qualities (Borsky *et al.*, 2017). While audio features consistently outperformed EGG, with accuracy reaching nearly 80% for RF and SVM, classification based solely on EGG averaged around 55–57%. Combining

EGG with other modalities did not significantly improve performance. The authors found that the COVAREP feature set (including glottal source features and harmonic model features) performed best, however, these results were only obtained from audio signals alone – no COVAREP were tested on EGG signals. The application of MFCC features in classification of breathy, modal, and strained voice also delivered accuracy between 74 and 79%, nonetheless, it was far more successful for audio than for EGG.

Despite advancements, EGG-based classification systems face challenges, including variability in signal quality and limited accuracy for complex voice qualities.

3.2.2 Deep Learning Approach

Recent advancements in deep learning have significantly improved the accuracy of EGG-based laryngeal pathology classification systems. One of the earliest examples was proposed by Ritchings *et al.* (1999), who found that an EGG signal and its derivative parameters (referred to as “short-term” parameters, including harmonic linearity measure, glottal noise and the Gaussian distribution calculated from positions of the first five harmonics) can be fed into a Multi-Layer Perceptron (MLP) model, achieving an accuracy of 80% in detection of pathological laryges. In 2001, the authors extended their previous work developing a system for objective assessment of voice quality in cancer patients. By combining the short-term parameters and long-term features (e.g., mean and standard deviation of fundamental frequency and voiced signal percentage), their system – trained with a back-propagation algorithm – achieved 92% accuracy in a 7-grade classification scheme. The work emphasised the importance of all parameters in the process of classification; the system increased its performance accuracy from 26.5% with just one parameter of the first harmonic’s Gaussian distribution, to 92% with the application of all short- and long-term features.

Considering deep learning models tend to perform significantly better provided with large datasets, most accurate systems utilising advanced neural network architectures emerged recently, as more public databases became accessible. Muhammad and Alhussain (2021) utilised pre-trained Convolutional Neural Networks (CNNs) including ResNet50, Xception, and MobileNet to extract features from laryngeal bioimpedance signals (and the spectrograms derived from them) obtained from SVD. Features were subsequently fed into a BiLSTM model, achieving 95.65% accuracy. Similarly, Islam *et al.* (2022) implemented a two-stage CNN system to classify dysphonia, laryngitis, and vocal fold polyps using raw audio and EGG signals obtained from SVD – first, the healthy and pathological signals were classified in a binary approach, leaving pathological signals for multi-class discrimination in the second CNN stage. The proposed approach achieved an average accuracy of 73.33% for EGG signals and 82.34% for audio in binary classification, and below 80% in multi-class; while audio outperformed EGG in distinguishing healthy from pathological voices, EGG showed superior performance in classifying specific pathologies.

In a separate work, Islam *et al.* proposed a CNN-based laryngeal pathology detection system utilising MFCCs derived recordings of sustained vowel phonation obtained from SVD. The proposed system reached the accuracy of 50.41% for EGG signals (58.33% for healthy, 42.50% accuracy for detection of pathological voices), and 74.28% for audio (73.33% for healthy, 75.00% for detection of pathological voices). These results suggest that MFCCs lower the ability of deep learning models to capture features related to specific pathologies, therefore, it is reasonable to investigate more appropriate feature extraction methods for laryngeal bioimpedance signals.

The potential of multimodal systems has also been explored. Miliaresi *et al.* (2022) integrated all three modalities offered by SVD; audio, EGG wavegrams, and demographic data. The achieved accuracy was 89.30% for classification between dysphonia, laryngitis,

and vocal fold paralysis in a middle (or hybrid) fusion multimodal approach (Gadzicki *et al.*, 2020) – processing of each modality through a different branch of deep learning model to eventually concatenate all branches in a fully connected layer. While tested on EGG signals alone, the system yielded 59.40% accuracy for EGG wavegrams, and 26.50% for EGG spectrograms. Notably, EGG wavegrams – derived following the methodology proposed by Herbst *et al.* (2010) – retained more classification-relevant features compared to spectrograms.

The study by Geng *et al.* (2022) can be closely compared to the work of Miliaresi *et al.* (2022), as both employed multimodal approaches using signals from SVD. Among the studies discussed, this work covered the broadest range of vocal tract pathologies, including leukoplakia, laryngitis, Reinke's oedema, vocal fold paralysis, nodules, and polyps. This study relied on CNN-based models, with incorporated Multimodal Transfer Module, distinguishing this approach to that of Miliaresi *et al.*'s. Additionally, instead of using MFCCs and wavegrams, the audio and EGG signals in Geng's work were processed as Mel-spectrograms. The proposed model achieved 100% accuracy in binary classification of healthy versus pathological signals, and 98.02% accuracy for multi-class pathology classification. The impressive performance of this model may be partially attributed to the utilisation of the pre-trained ResNet18 CNN model (Geng at al., 2022).

Further research by Kumar *et al.* (2023) investigated feature extraction methods for EGG signal classification; 25 various feature extraction algorithms were evaluated using four classifiers: support vector machine (SVM), k-nearest neighbour (KNN), ensemble learner (EL) and deep neural network (DNN). By applying the minimum redundancy maximum relevance (MRMR) algorithm, they identified the ERB spectrum features and GTCCs as the most informative, achieving a maximum accuracy of 93.15% using an Ensemble Learner.

Most recently, work in laryngeal imaging has adopted self-supervised methods, for instance, autoencoders. Darvish and Kist (2024) proposed a variational autoencoder (VAE) framework to synthesise realistic glottal area waveforms, offering a data augmentation strategy for addressing the lack of labelled laryngeal datasets. Building upon the publicly available Laryngoscope8 dataset (Yin et al., 2021), Yan et al. (2023) introduced a dual-transformer architecture tailored for medical image classification, which achieved 85.2% accuracy on the Laryngoscope8 test set, establishing a strong state-of-the-art baseline in image-based laryngeal disease classification. Subsequently, Xu et al. (2023) advanced this line of research by proposing an automatic annotation pipeline for laryngeal images. Their method was based on the Segment Anything Model (SAM), which combined a vision transformer pre-trained with a masked autoencoder (MAE) for image encoding, alongside a prompt encoder and a masked decoder that predicted segmentation masks. This approach facilitated efficient annotation and classification within the same dataset, chieving 79% accuracy.

While these works demonstrate the promise of transformer-based and self-supervised methods for laryngeal dataset enrichment, it is important to note that they operate exclusively on endoscopic images. In this study, we aim to explore diagnosis directly from physiological signals such as voice recordings and bioimpedance waveforms, offering a complementary pathway that avoids reliance on additional image processing.

Deep learning has shown significant potential in EGG-based pathology classification, particularly when supported by robust datasets, appropriate feature extraction methods, and multimodal integration. Recent studies highlight the emergence of the ERB spectrum-derived features as especially effective for laryngeal bioimpedance signals, offering high relevance for accurate classification. However, further research is needed to refine classification performance of complex multi-class pathologies relying solely on

bioimpedance measurements. Also, further investigation into application of continuous speech rather than sustained phonation needs to be performed.

3.2.3 Summary of Laryngeal Bioimpedance-based Methods

Electroglottography (EGG) has emerged as a valuable tool for laryngeal pathology examination due to its non-invasive nature and ability to capture glottal dynamics. Early systems relied on statistical methods and statistical measures (Childers and Bae, 1992; Jiang *et al.*, 1998; Hosokawa *et al.*, 2014). While these methods demonstrated promise, especially for detection of glottal instances (Thomas and Naylor, 2009; Deshpande and Manikandan, 2017), the reliability of EGG signals was questioned (Jiang *et al.*, 1998.; Borsky *et al.*, 2017). The variability in laryngeal bioimpedance signal quality and limited generalisability across multi-class classification highlighted the need for more advanced approaches.

Deep learning approaches have significantly enhanced classification accuracy by utilising larger datasets and advanced model architectures; for instance, CNNs, BiLSTMs, as well as autoencoders for automatic data annotation and classification (Yan et al., 2023; Xu et al., 2023), as well as augmentation (Darvish and Kist, 2024). These systems have successfully utilised EGG signals in various forms, including spectrograms and wavegrams, and often incorporate multimodal frameworks combining audio and demographic data (Ritchings *et al.*, 2001; Muhammad and Alhussain, 2021; Islam *et al.*, 2022; Miliaresi *et al.*, 2022; Geng *et al.*, 2022). Multimodal systems, particularly those utilising transfer learning and robust preprocessing, have demonstrated substantial improvements in multi-class classification tasks (Geng *et al.*, 2022).

Recent research has identified the ERB spectrum-derived features as particularly effective for EGG-based classification (Kumar *et al.*, 2023). Despite these advancements,

further exploration of robust feature extraction techniques and multimodal integration is essential to realise full clinical potential of laryngeal bioimpedance.

The choice of phonation types (sustained phonation versus short yet continuous speech utterances) play a crucial role in the accuracy of laryngeal pathology classifiers, alongside spectral representations and feature extraction methods. Nearly all research investigated the data collected from participants during sustained vowel phonation (Childers and Bae, 1992; Jiang *et al.*, 1998; Rosa *et al.*, 1999; Ritchings *et al.*, 2001; Borsky *et al.*, 2017; Nacci *et al.*, 2020; Muhammad and Alhussain, 2021; Miliaresi *et al.*, 2022; Islam *et al.*, 2022; Geng *et al.*, 2022). According to literature, the reasons for preference of sustained phonation over speech are stable positioning of the epiglottis, consistent fundamental frequency, and absence of complex articulatory movements (Jiang *et al.*, 1998; Islam *et al.*, 2022). However, the dynamic glottal changes that occur during continuous speech may offer additional insights into pathological patterns in bioimpedance signals.

Table 3.2 provides a summary of selected previous work on classification of laryngeal pathologies based on laryngeal bioimpedance signals.

Table 3.2: Summary of seminal published work on laryngeal pathology classification based on laryngeal bioimpedance.

REFERENCE	DATASET	OBJECTIVE	METHODS	FINDINGS
Childers and Bae, 1992.	81 speakers (52 control, 29 pathological). Male and Female.	Binary classification (control vs pathological).	DATA: Custom dataset – Audio and Glottal Bioimpedance (EGG) – sustained vowel /i/ for 2 s. FEATURES: Linear Prediction and Vector Quantisation (audio), Perturbation Analysis (bioimpedance). CLASSIFICATION: Closed-threshold test and discriminant analysis.	SENSITIVITY: Audio Signals: 75.9% for pitch asynchronous LPC analysis method, 44.8% with pitch synchronous analysis method. Bioimpedance: 69.0% (manual counting of features whose values exceeded the threshold).
Jiang *et al.*, 1998.	36 speakers (7 control, 10 recurrent paralysis, 9 superior paralysis in training dataset; 7 control, 22 recurrent paralysis, 7 superior paralysis). Male and Female.	Binary classification (control vs paralysis).	DATA: Custom dataset – Glottal Bioimpedance (EGG) and photoglottogram (PGG) – sustained vowel /i/ for 500 ms. FEATURES: EGG and PGG derivatives (speed quotient, open quotient, closed quotient).	SENSITIVITY: Audio integrated with bioimpedance: 43% accuracy for control detection, 73% for recurrent paralysis, 57% for superior paralysis. ACCURACY:

			CLASSIFICATION: Probability between new signals and knowledge database (comparing area of overlap between distribution of derivative in signals).	64% accuracy for all testing dataset samples.
Ritchings *et al.*, 2001.	77 pathological (at different stages of cancer treatment). Male only.	Multi-class classification (stages of larynx cancer treatment and recovery).	DATA: Custom dataset – Glottal Bioimpedance (EGG) – sustained vowel /i/ for 3 s. FEATURES: Mean fundamental frequency, its standard deviation, voiced signal percentage, harmonic linearity measure, glottal noise, Gaussian distribution of position of first five harmonics. CLASSIFICATION: 2-layer 7-output Multi-layer Perceptron (MLP) neural networks.	SENSITIVITY: 92%
Borsky *et al.*, 2017.	28 control (unaffected by vocal tract pathologies). Male and Female.	Multi-class classification (four voice types of modal, breathy, strained, and rough).	DATA: Custom dataset – Audio, Glottal Inverse Filtered Signal and Glottal Bioimpedance (EGG) – sustained vowel /a/, /e/, /i/, /o/, /u/ for 2-5 s. FEATURES: COVAREP feature set (audio), MFCCs and first order dynamic MFCCs (audio, glottal inverse filtered signal, bioimpedance). CLASSIFICATION: Gaussian mixture model (GMM), Support Vector Machine (SVM), Random Forest classifier (RF), Deep Neural Network (DNN).	ACCURACY: Bioimpedance: 56% Audio integrated with bioimpedance: 78.38% for SVM, 79.53% for RF, 56.26% for GMM, 76.7% for DNN.
Nacci *et al.*, 2020.	125 speakers (36 control, 24 functional dysphonia, 21 bilateral vocal polyps, 23 unilateral polyps, 21 unilateral cysts). Male and Female.	Multi-class assessment (functional dysphonia and organic pathologies – polyps, cyst, nodules).	DATA: Custom dataset – Glottal Bioimpedance (EGG) – sustained vowel /a/. FEATURES: Amplitude-speed combined analysis expressed as Variability Index (VI), Kruskal-Wallis test. CLASSIFICATION: Mann-Whitney U test corrected with Bonferroni.	SPECIFICITY: 66.7% for VI-tot and 77.8% for VI-Q2%. Mann-Whitney U test corrected with Bonferroni: P Values for comparisons between functional dysphonia, polyps, nodules, cysts: <0.021.
Muhammad and Alhussain, 2021.	1072 speakers (281 control, 791 pathological). Male and Female.	Binary classification (control vs pathological).	DATA: Saarbruecken Voice Database – Audio and Glottal Bioimpedance (EGG) – sustained vowel /a/ for 1-3 s. FEATURE: Spectrograms and Mel-spectrograms, Pre-trained Convolutional Neural Network (ResNet50, Xception, and MobileNet models tested). CLASSIFICATION: Bi-directional Long Short-term Memory Network.	ACCURACY: Audio: 93.94% Bioimpedance: 93.71% Audio integrated with bioimpedance: 95.65%
Miliaresi *et al.*, 2022.	1241 speakers (687 control, 140 laryngitis, 204 dysphonia, 210 paralysis). Male and Female.	Multi-class classification (control vs dysphonia vs laryngitis vs paralysis).	DATA: Saarbruecken Voice Database – Audio, Glottal Bioimpedance (EGG) – sustained vowel /a/ for 0.5-3 s, medical data. FEATURES:	ACCURACY: Bioimpedance: 58.30% for closed quotient only, 59.40% for bioimpedance wavegrams only, 26.50% for bioimpedance spectrograms only.

			Mel Filter banks, MFCC, MFCC derivatives, jitter, fundamental frequency, and harmonic to noise ratio (audio). Wavegrams, spectrograms and closed quotient (bioimpedance). CLASSIFICATION: Modular deep learning using Feed Forward Neural Network and Convolutional Neural Network.	Audio integrated with bioimpedance: 81.10% for acoustic signal and closed quotient, 82.60% for acoustic signal and bioimpedance wavegrams. Integrated all modalities: 89.30%.
Islam *et al.*, 2022.	215 speakers (150 control, 30 dysphonia, 25 laryngitis, 10 vocal fold nodules). Male and Female.	Binary classification (control vs pathological) AND multi-class classification (laryngitis vs polyp vs dysphonia).	DATA: Saarbruecken Voice Database – Audio and Glottal Bioimpedance (EGG) – sustained vowel /a/. FEATURES: Feature extraction using Convolutional Neural Network. CLASSIFICATION: Dual cascaded Convolutional Neural Networks.	ACCURACY: Audio: Binary classification: 80.30%, Multi-class: 76.48%. Bioimpedance: Binary classification: 72.10%, Multi-class: 88.67%.
Islam *et al.*, 2022.	50 speakers (25 healthy, 25 dysphonia). Male and Female.	Binary classification (control vs pathological).	DATA: Saarbruecken Voice Database – Audio and Glottal Bioimpedance (EGG) – sustained vowel /a/. FEATURES: MFCC (audio, bioimpedance). CLASSIFICATION: Convolutional Neural Networks.	ACCURACY: Audio: 74.28%. Bioimpedance: 50.41%.
Geng *et al.*, 2022.	1179 speakers (613 control, 566 pathological – leukoplakia, laryngitis, Reinke's Oedema, paralysis, vocal nodules and polyps). Male and Female.	Binary classification (control vs pathological) AND multi-class classification (leukoplakia, laryngitis, Reinke's Oedema, paralysis, vocal nodules and polyps).	DATA: Saarbruecken Voice Database – Audio and Glottal Bioimpedance (EGG) – sustained vowel /a/. FEATURES: Mel-spectrograms (audio and bioimpedance). CLASSIFICATION: Convolutional Neural Network (pre-trained ResNet18 model) with multimodal transfer module.	ACCURACY: Audio integrated with bioimpedance: Binary classification: 100%. Multi-class: ACCURACY: 98.02%, RECALL: 98.23%, SPECIFICITY: 97.82% F1 SCORE: 97.95%.
Kumar *et al.*, 2023.	606 speakers (303 control, 303 pathological – Reinke's oedema, vocal fold polyp, leukoplakia, and dysphonia). Male and Female.	Binary classification (control vs pathological).	DATA: Saarbruecken Voice Database – Audio and Glottal Bioimpedance (EGG) – sustained vowel /a/, /i/, and /u/ for 25ms. FEATURES: Various methods assessed with minimum redundancy maximum relevance algorithm (MRMR), including MFCC, Equivalent Rectangular Bandwidth (ERB) Spectrum and Gammatone Cepstral Coefficients (GTCC). CLASSIFICATION: Support vector machine (SVM), k-nearest neighbour (KNN), Ensemble Learner and Neural Networks (NN).	Bioimpedance: Best performing – Ensemble Learner on ERB and GTCC: ACCURACY: 93.15% PRECISION: 96.70%, RECALL: 90.29%, F1 SCORE: 93.38%. Audio integrated with bioimpedance: ACCURACY: 79.97%.

3.3. SUMMARY

The above chapter reviews the existing literature on laryngeal pathology classification systems based on two data modalities – audio, and laryngeal bioimpedance (EGG).

The existing body of research on laryngeal pathology classification has primarily relied on single-modality approaches, majorly focusing on the investigation of sustained phonation signals. It is an assumption commonly reappearing in the literature, that sustained phonation is superior to continuous speech in detection and evaluation of pathological patterns in digital signals (Rosa *et al.*, 1999; Henríquez *et al.*, 2009; Markaki and Stylianou, 2011; Islam *et al.*, 2022). Based on very limited research available (Wang *et al.*, 2022), this belief may be a misconception. It is, however, crucial to investigate the performance of sustained vowel phonation against continuous speech using one classification method to thoroughly assess and evaluate the hypothetical superiority of one phonation type over another in conveying pathological patterns.

Audio-based approaches, utilising statistical and machine learning methods, have demonstrated high accuracy in binary classification tasks using various features including pitch and amplitude perturbation, harmonic-to-noise ratio, correlation entropy, and modulation spectral features (Peng *et al.*, 2007; Henríquez *et al.*, 2009; Markaki and Stylianou, 2011). However, their performance declines in multi-class discrimination (Borsky *et al.*, 2017; Islam *et al.*, 2022).

The laryngeal bioimpedance provides a physiological perspective on laryngeal function and has evolved from statistical parameter extraction to deep learning-based classification (Childers and Bae, 1992; Kumar *et al.*, 2023). However, literature suggests that EGG alone does not offer a consistently reliable means of laryngeal pathology detection. Traditional statistical approaches, such as glottal cycle analysis and frequency-domain features, have been used for classification but often yield lower accuracies than audio-based methods

(Borsky *et al.*, 2017; Miliaresi *et al.*, 2022; Islam *et al.*, 2022). Even with modern deep learning methods such as CNN, the accuracy remains dataset-dependent and lacks generalisability across diverse patient populations (Islam *et al.*, 2022; Kumar *et al.*, 2023). Furthermore, inconsistencies in EGG waveform interpretation, susceptibility to artifacts, and variability in electrode placement further limit the reliability of EGG as a standalone diagnostic tool (Baken, 1992; Herbst, 2019).

Given that audio alone does not yield sufficient accuracy for multi-class discrimination and that EGG lacks reliability as a standalone classification modality, the viable solution is a multimodal approach that combines both types of signals. Recent studies have demonstrated that integrating these two modalities significantly improves the accuracy of proposed laryngeal pathology classification systems (Miliaresi *et al.*, 2022; Muhammad and Alhussain, 2021; Geng *et al.*, 2022).

Following the conclusions drawn from this review of existing research, in this work we demonstrate that a multimodal deep learning framework substantially enhances the performance of laryngeal pathology classification. Furthermore, we focus on delivering the comparative results for both phonation types – continuous speech and sustained phonation – demonstrating that continuous speech delivers better results in classification of pathological signals. Through comparative analyses, we establish that multimodal systems consistently outperform unimodal approaches, offering a robust, more accurate, and clinically viable methodology for automated pathology detection and classification.

Chapter 4

Datasets and Initial Data Analysis

This chapter outlines the datasets used in this study and the analytical steps undertaken to ensure their suitability for laryngeal pathology classification tasks. In this chapter, we provide the detailed description of the data collection process and its demographic analysis, as well as the applied data preprocessing methods and the final dataset setup assembled based upon the medical relevance and computational feasibility. For the purposes of completeness and generalisability, the research uses two datasets: a custom dataset collected specifically for the purposes of this study, and the Saarbruecken Voice Database – SVD (Pützer and Koreman, 1997; Saarbruecken Voice Database: Handbook, 2023). Both datasets undergo the same data preprocessing pipeline including a rigorous shuffling of participants to ensure unbiased data distribution.

Additionally, an extensive exploratory data analysis (EDA) is performed on the custom dataset to assess the classification potential of the gathered samples, as well as the medical and computational feasibility of laryngeal pathology classification based on audio and laryngeal bioimpedance recordings.

The chapter is divided into four main sections. The first section (*4.1. Custom Dataset*) provides a detailed description of the data collection process, the demographic distribution of participants recruited for the collection of custom data, and the data preprocessing pipeline applied later to both the custom dataset as well as SVD. The second section (*4.2. Preliminary Investigation of Custom Dataset Classification*) focuses on initial classification experiments investigating the computational feasibility of developing the intended laryngeal pathology classification system and its medical relevance for diagnostics. This section highlights the limitations of the original pathology grouping, revealing the impact of speaker-

dependent biases on the accuracy of the classification system, and justifying the reorganisation of pathology classes into the three broad categories: cancerous and precancerous lesions, neuromuscular disorders, and healthy cases.

In the third section (*4.3. Exploratory Data Analysis*) we perform the exploratory analysis of the custom dataset to identify and analyse the main characteristics of each investigated laryngeal condition group. The EDA examines the statistical parameters calculated for each subgroup the final dataset, assessing the presence of potential clustering tendencies and evaluating the class separability before building the intended classification system. For that, Principal Component Analysis (PCA) is applied to different feature sets derived from speech and sustained phonation data of both modalities – audio and laryngeal bioimpedance. Subsequently, we assess the separability of the data clusters using Hopkins statistics and Euclidean distances. Furthermore, a primary investigation of the multimodal approaches to data classification are evaluated. The performed EDA provides critical insights into the diagnostic relevance of the data and highlights the limitations of the statistical analysis, reinforcing the need for more sophisticated methods for the intended classification of laryngeal pathologies – for instance, the deep learning.

In the final section (*4.4. Saarbruecken Voice Database*), we evaluate the SVD database and its suitability for the intended laryngeal pathology classification system. This includes an examination of dataset limitations, a careful participant selection process, and the alignment of SVD data preprocessing with the same refined pipeline as that applied to the custom dataset.

In summary, this chapter establishes the foundation for data-driven classification of laryngeal pathologies by investigating the data class separability. The conclusions drawn from this chapter are used further in this research to inform the machine learning and deep learning methodologies developed in subsequent stages of this study.

4. DATASETS AND INITIAL DATA ANALYSIS

4.1. CUSTOM DATASET

This research on computational methods for laryngeal pathology classification relies on two categories of data: audio and laryngeal bioimpedance. Both data modalities were collected simultaneously, ensuring all audio recordings have their counterparts in the form of bioimpedance measurements.

The data was collected from participants affected by one of the laryngeal pathologies described in the *2.4. Investigated Laryngeal Pathologies* section: malignant growths of the vocal fold area (cancerous and precancerous growths), other growths outside of the vocal fold area (benign growths), neuromuscular disorders (predominantly vocal fold paralysis), laryngitis, Reinke's Oedema, and functional dysphonia. The audio and bioimpedance signals were also gathered from a group of participants unaffected by any of the laryngeal pathologies to form a control group dataset. One sample of audio data and the corresponding bioimpedance signal from each category have been randomly selected and presented as figures in the *2.4. Investigated Laryngeal Pathologies* section of this document. The figures show time-domain representations of each data modality (the waveforms), as well as their corresponding spectral content (frequency domain).

The collection of all relevant data was possible due to the collaboration we had established with ENT doctors from Czerniakowski Hospital, Warsaw, Poland. All recordings were carried out in accordance with *General Data Protection Regulation, outlined in the Regulation (EU) 2016/679* (General Data Protection Regulation). The collected data is anonymous and classified exclusively regarding its context. Upon completion of data collection from each participant, the individual recordings were offered available to them upon request.

The following sections detail the creation of the custom dataset developed for this study:

- *4.1.1 Data Collection*: The overview of the data collection process, with an emphasis on the recorded phonation types and the equipment used.
- *4.1.2 Participants and Demographic Data Analysis*: The information on participants and demographic analysis of the dataset.
- *4.1.3 Data Preprocessing*: The overview of the methods used for data preprocessing.
- *4.1.4 Data Split and Categorisation*: The process of organising data into corresponding classes.

4.1.1 Data Collection

All data, including both the control group as well as pathological subset, was collected in Czerniakowski Hospital, Warsaw, Poland. A total of 156 participants were recruited: 136 affected by a laryngeal pathology, and 20 healthy control participants.

During the data collection process, the participants were asked to perform the following tasks:

- breathing in and out through the nose,
- breathing in and out through the mouth,
- coughing,
- sustaining vowels /a/, /e/, /i/, /o/, and /u/, each repeated at pitches C1, F1, and A1,
- sustaining “mormorando” sound (murmuring with a closed mouth),
- reading one paragraph of text in Polish language.

The paragraph read: “The above study aims to develop an innovative classification system for respiratory abnormalities in the form of a computer application. The intended system is to establish a potential diagnosis of respiratory disorders by evaluating recordings of the patient’s voice while breathing, speaking, coughing, and singing.”

Since all participants were Polish speaking, the paragraph of text had been translated to Polish language, as follows: "*Powyższe badanie ma na celu opracowanie nowatorskiego systemu klasyfikacji nieprawidłowości oddechowych w formie aplikacji komputerowej. Zamierzony system ma umożliwić ustalenie potencjalnej diagnozy zaburzeń w układzie oddechowym poprzez ewaluację nagrań głosu pacjenta podczas oddychania, mówienia, kaszlu i śpiewu. System ma opierać się na podejściu audiometrycznym (ewaluacja nagrań audio) z jednoczesnym pomiarem bioimpedancji krtani*".

The process of data collection took place in two stages:

- Stage 1 – Pathological Data Collection:
    - conducted between November 2022 and March 2023,
    - involved collecting data from participants with the diagnosis of laryngeal pathology.
- Stage 2 – Control Data Collection:
    - conducted in September 2023,
    - involved collecting control data from healthy participants (those with no diagnosis of laryngeal pathology).

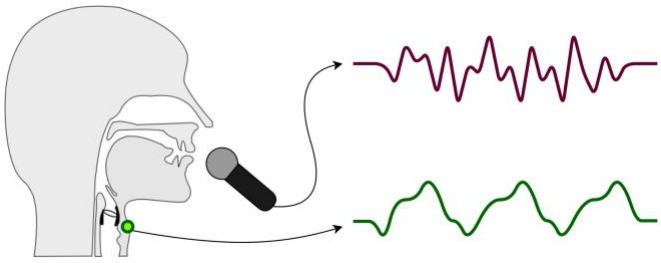

Figure 4.1: Schematic representation of data collection process.

Both data modalities – audio and bioimpedance – were recorded simultaneously using Scarlett 2i4 audio interface and Logic Pro X digital audio workstation (DAW). The signals were obtained "raw" – during the recording stage, no compression or other preprocessing was applied.

For the acquisition of audio data, a dynamic SM57 microphone was used. The bioimpedance measurements were obtained using a Kay Model 6103 electroglottograph, the most widely used EGG in laryngeal pathology research (Nacci, 2020; Hosokawa *et al.*, 2014). The two electrodes were placed bilaterally at the level of trachea at the wings of the thyroid cartilage, corresponding to the region where the vocal cords reside, and they were secured with Velcro tape. To ensure the correct placement of the electrodes, the signal was continuously monitored using PicoScope PC Ocilloscope.

After the initial segmentation of the data depending on the content of each signal (further described in detail in the *4.1.4 Data Split and Categorisation* section), all recordings were exported in a form of mono WAV files with the sampling rate of 44100 Hz and the bit depth of 16 bits per sample.

The data collection procedure resulted in 1469 recordings of bioimpedance measurements and 1469 recordings of audio events, both gathered from the total of 136 participants affected by laryngeal pathologies, as well as 392 recordings of bioimpedance and 392 recordings of audio gathered from 20 control participants. In total, this resulted in 1861 recordings for each data modality.

4.1.2 Participants and Demographic Data Analysis

Due to the collaboration established with Czerniakowski Hospital based in Poland, Warsaw, all recruited participants were of Polish nationality. To minimise variability due to accent or pronunciation, all control group data was also collected exclusively from Polish-speaking participants under identical conditions and using the same equipment as for the pathological group.

A total of 156 participants were recruited; 20 control subjects experiencing no impairments within the phonatory tract (no diagnosis any laryngeal pathology), and 136

subjects that were affected by one of laryngeal pathologies listed in section *2.4. Investigated Laryngeal Pathologies*. All participants were assessed by a clinician specialised in phoniatrics immediately before the beginning of data collection process to ensure accurate diagnoses. These assessments consisted of auditory evaluation of a subject's voice, as well as an endoscopic evaluation using a laryngoscope with a stroboscopic light.

Due to technical problems with the recordings, or their poor quality caused by the immense struggle of a patient during the recording process, the data collected from 15 participants from the pathological group were removed from the overall dataset and suspended from further processing. Figure 4.2 depicts the final number of participants in each classification group, including all pathologies and control group participants as a separate class.

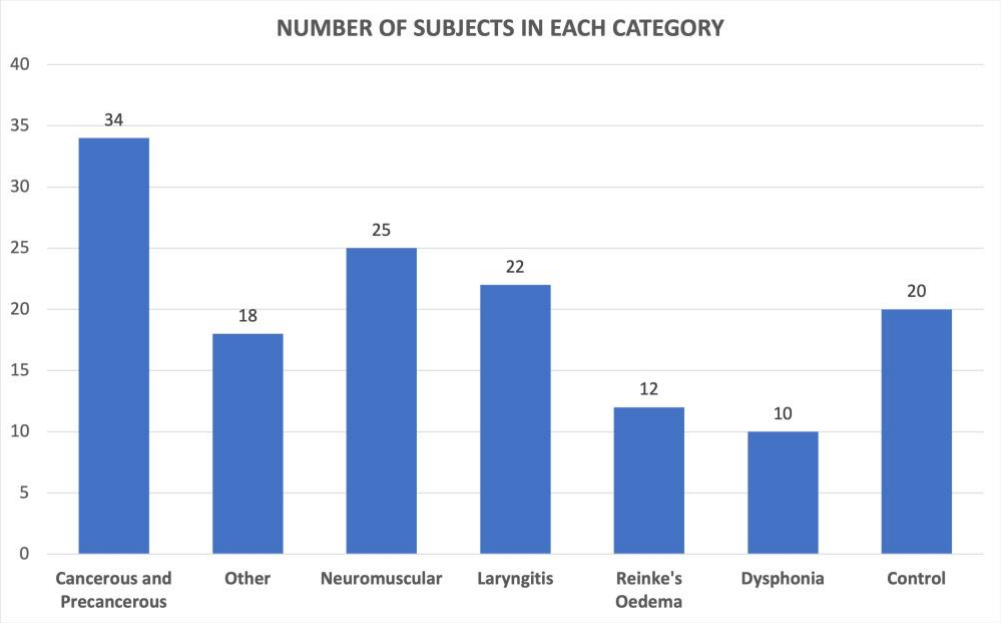

Figure 4.2: Number of participants in each category.

The control group dataset was gathered from 20 participants, including 11 females and 9 male participants. All subjects from the control group were between 26 and 56 years old, with the average age of 39 (SD of 11.54) and median of 37.

Among the 136 participants affected by laryngeal pathologies, 79 were female and 57 were male. The age range in the pathological group was also significantly broader to that of the control group, ranging from 21 years of age to 86, with an average of 57 (SD of 15.12) and median of 59. Table 4.1 shows mean, median, and standard deviation of age for males and females from both control and pathological groups, respectively.

Table 4.1: Age representation of males and females from both control and pathological groups.

Age of Participants	**Range**	**Average**	**SD**	**Median**
Control – females	26-56	40	11.48	38.5
Control – males	26-55	38	12.51	33
Pathological – females	21-86	56	14.72	57
Pathological – males	29-86	59	15.61	61

Certain laryngeal pathologies demonstrated gender-related prevalence. An example of such disorder is the vocal fold paralysis. The laryngeal nerve paralysis is often triggered by a thyroid gland tumour or thyroid related trauma (for instance, a surgery) due to the proximity of a thyroid and laryngeal nerves (Myssiorek, 2004). According to literature, thyroid-related conditions are four times more common in women than men (Vanderpump, 2011), which directly correlates with the incidence of vocal cord paralysis in women.

Conversely, the malignant growths of vocal folds are more commonly observed in male subjects (Altman, 2007). According to the American Cancer Society (2023), men are over four times more likely to develop laryngeal cancer than women, with risks estimated at 1 in 190 for men and 1 in 830 for women. In 60% of cases, the laryngeal cancer begins with growths on the glottal or vocal fold area.

The following figure (Figure 4.3) shows the number of males and females in each pathological category. The subsequent figure (Figure 4.4) shows how the frequency of the chosen laryngeal pathologies changes with age for both genders.

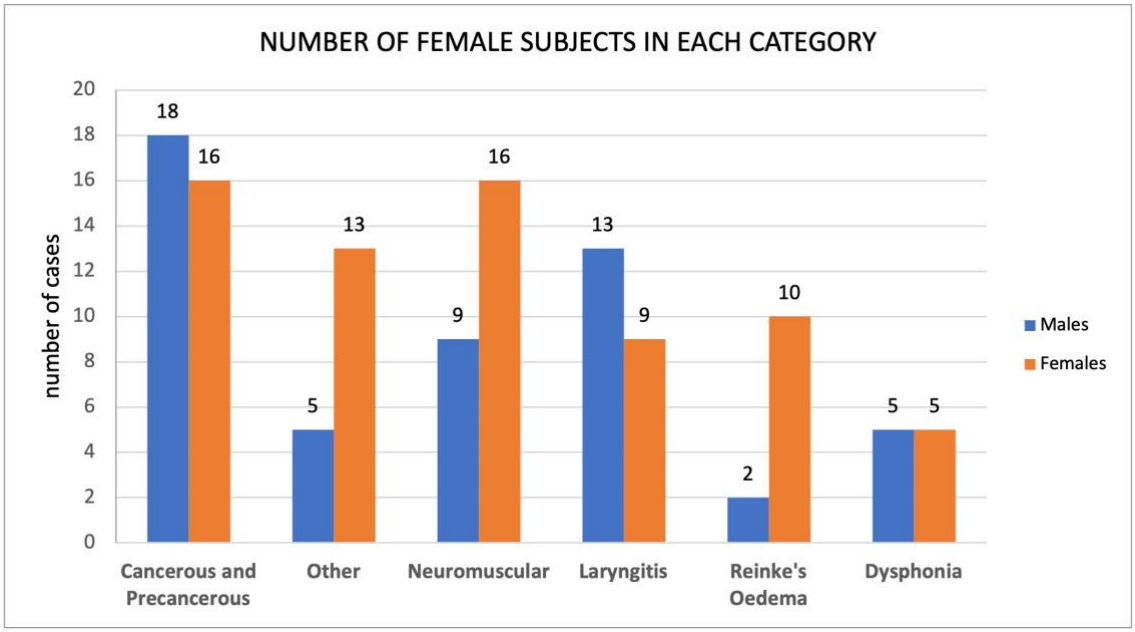

Figure 4.3: Number of males and females in each pathological category.

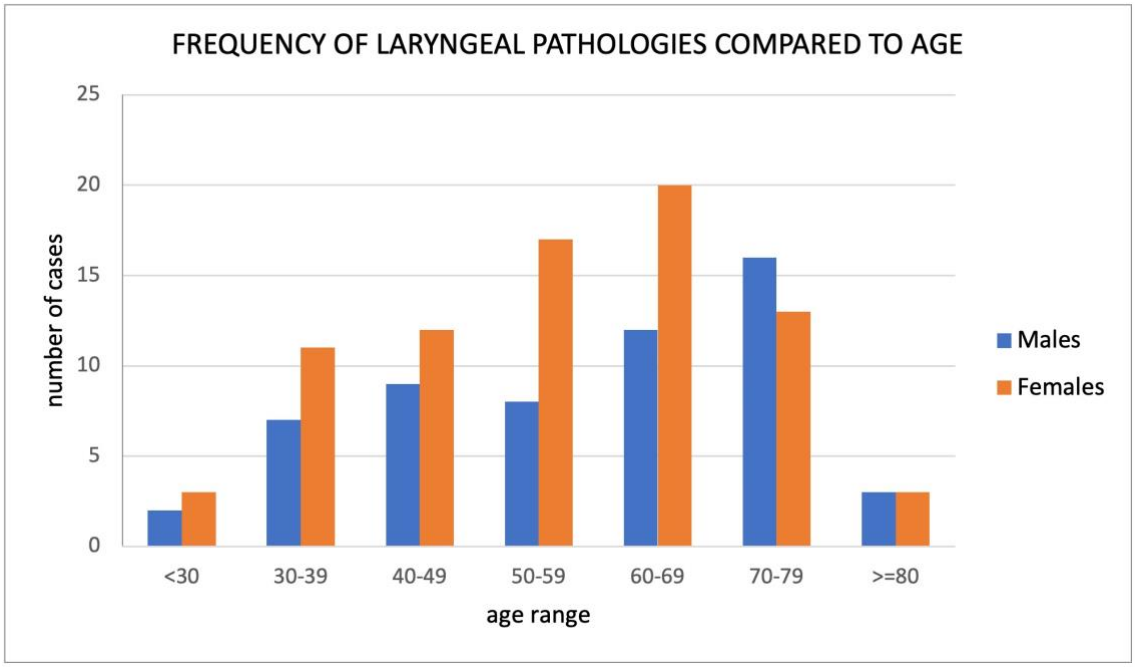

Figure 4.4: Frequency of laryngeal pathology occurrence compared to the age group.

Reinke's Oedema is another example of a laryngeal pathology that is more commonly diagnosed in females than males (Dewan *et al.*, 2022). Reinke's Oedema, a persistent respiratory condition, predominantly manifests in adults who have a history of significant exposure to tobacco smoke, vocal misuse, or laryngopharyngeal reflux (for instance, gastroesophageal reflux disease – GERD). It leads to abnormal swelling and the development of mass on the vocal folds, resulting in impaired pitch control, typically leading to a deeper voice quality and increased vocal roughness. Due to the more sudden and audible change of voice, women are more inclined to seek the appropriate diagnosis, as the distinctive pitch alteration tends to be more conspicuous in females compared to males (Dewan *et al.*, 2022). This principal applies to most laryngeal pathologies, thereby explaining why more women than men receive a related diagnosis.

Furthermore, certain laryngeal pathologies can be associated with different age ranges. The following figure represents the frequency of individual diseases present in the collected dataset arranged by age groups (Figure 4.5).

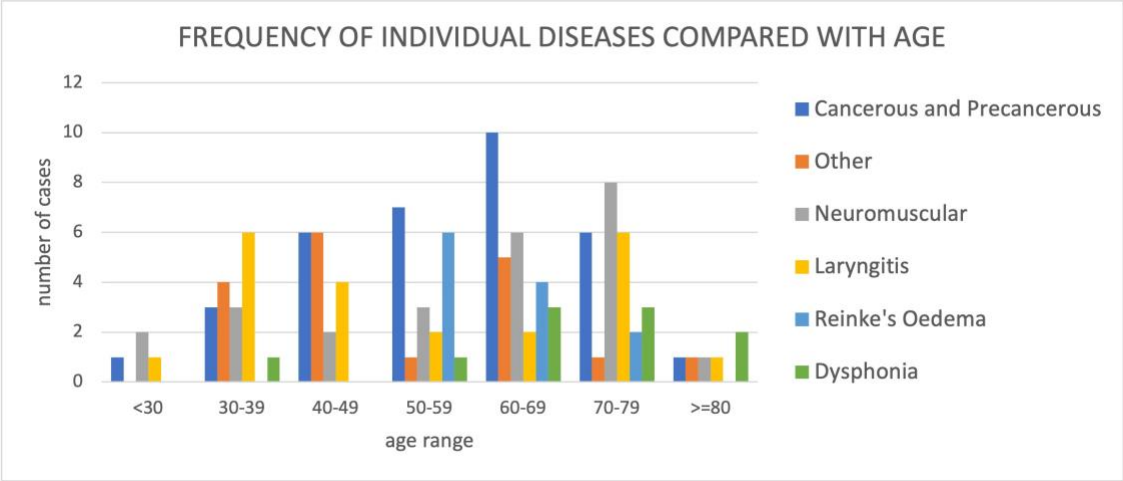

Figure 4.5: Frequency of occurrence of individual laryngeal pathologies compared to the age group.

The larynx experiences numerous physiological and structural changes associated with aging. Those can include muscles atrophy or thinning of laryngeal ligament that often lead to dysphonia (Gregory *et al.*, 2012). Naturally, the probability of diagnosing functional

dysphonia that is not related to any physical or psychological trauma increases with patient's age. In contrast, the incidence of inflammatory conditions such as laryngitis tends to decline between the ages of 50 and 70, possibly due to reduced occupational exposure to pathogens.

The initial analysis of the collected data provided valuable insights into the distribution of pathologies, which aligns with medical literature available on the matter, and informed the selection of appropriate data, followed by preprocessing techniques for further investigation.

4.1.3 Data Preprocessing

The data preprocessing for this study was completed in two stages. The initial stage focused on preparing the dataset for further analysis by standardising it and ensuring its consistency across the modalities – the methods used for the initial preprocessing stage were identical for both audio and the bioimpedance signals. After completing this stage, the preliminary data classification testing was pursued (for details, see *4.2. Preliminary Investigation of Custom Dataset Classification*). The preliminary testing of classification methods performed on data subjected to only the initial stage of preprocessing revealed opportunities for further optimisation of the preprocessing pipeline, particularly for the laryngeal bioimpedance signals. This led to the implementation of the latter stage of preprocessing, where key adjustments were made to enhance computational efficiency while preserving the diagnostic features most relevant for the appropriate laryngeal pathology classification.

4.1.3.1. Initial stage

The first stage of preprocessing was applied to the entire dataset, including both the bioimpedance and the audio recordings. The preprocessing, performed using MATLAB computing environment, consisted of three stages: band-pass filtering, peak normalisation,

and data pre-segmentation. All algorithms applied during the data preprocessing can be found in the algorithms folder created for the purposes of this research and are available upon request.

The band-pass filtering was a particularly important step for the preprocessing of bioimpedance measurements due to their acute voltage fluctuations. This step ensures the appropriate transformation of the Gx electroglottographic signal into stable Lx signal depicting vocal folds behaviour only (for further details on electroglottographic signals, see section *2.3. Human Phonation – Bioimpedance Signals (Electroglottography)*). However, there is a significant lack of comprehensive research identifying the specific frequency bands that are most relevant and representative for distinguishing various laryngeal pathologies, especially for laryngeal bioimpedance signals. To minimise the risk of excluding potentially important frequency components, we opted to apply minimal band-pass filtering. To preserve as much spectral information as possible while removing most frequency fluctuations unrelated to the phonatory tract health status, the 50-17000 Hz frequency band was chosen. The filtering was performed using two IIR filters with steepness of 0.95: a high-pass filter with a 50 Hz passband, and a low-pass filter with a 17000 Hz passband. All processed files were saved as WAV files in a separate folder.

The second stage of the initial data preprocessing focused on normalising all signals' amplitude. To preserve the real information contained within the recordings, and to avoid introducing any signal alterations, we chose to proceed with peak normalisation technique rather than compression. For that, the target peak level was set at -3 dB. The peak normalisation process consisted of the following steps: computation of the peak level of a processed signal, computation of the gain required to normalise that signal to the target level, the multiplication of that signal with the computed gain. All normalised files were saved to a separate folder.

Lastly, to standardise the recording length and to avoid signal zero-padding, all data was pre-segmented. Considering the variety of vocal exercises requested from the participants during the data collection, the length of the recordings varied from below 1 second, with majority between 2-7 seconds (the sustained vowel recordings), and maximum reaching a minute (the recordings of reading the text paragraphs). In table 4.2 we show the mean, SD, maximum, and minimum lengths of the recorded audio and bioimpedance signals.

Table 4.2: Statistics pertaining to the length of recorded data samples for both audio and bioimpedance.

	Mean	**SD**	**Min**	**Max**
Continuous Speech – All	26.17	7.25	13.50	60.00
Sustained Phonation – All	3.62	2.48	0.63	29.00
Continuous Speech – Pathological	26.67	7.45	13.50	60.00
Continuous Speech – Healthy	22.32	3.72	17.38	31.38
Sustained Phonation – Pathological	3.86	2.67	0.63	29.00
Sustained Phonation – Healthy	2.82	1.42	0.75	7.63

To address this variability, as well as to balance the representation of different classes (which was essential given the limited number of subjects), all recordings were divided into one-second-long segments, with each segment corresponding to 44100 samples. This process was completed using MATLAB computing environment by inputting an entire recording of a participant, separating the adjacent groups of 44100 samples within that recording and splitting them into separate arrays, subsequently saving each array of 44100 samples as a separate WAV file. By dividing longer segments, each segment could be treated as an independent sample without the need for zero-padding, thereby ensuring comparability across all modalities. The signal classification using full-length recordings was also tested, however, preliminary testing revealed that classification performance was

negatively impacted by zero-padding due to large variations in file lengths. Consequently, further testing on unsegmented recordings was discontinued.

4.1.3.2. Latter stage

Following the initial preprocessing, the preliminary data classification testing revealed areas for further optimisation, particularly for the bioimpedance signals. While the initial sampling rate of 44100 Hz was applied to both data modalities, it was observed that processing laryngeal bioimpedance signals at this rate introduced unnecessary computational complexity without yielding additional classification benefits (section *4.2. Preliminary Investigation of Custom Dataset Classification*). This can be attributed to the limited frequency content of the EGG signals. As illustrated in section *2.4: Investigated Laryngeal Pathologies* (see Figures 2.7-2.13 representing the spectral content over 5000 Hz present in both modalities), the frequency content of the bioimpedance data primarily constitutes the fundamental frequency and the first few harmonics, which may vary depending on the pathology. As such, a significantly lower sampling rate, such as 2048 Hz, was sufficient to capture the diagnostically relevant information.

To optimise the preprocessing pipeline, the second (“latter”) stage was established, introducing two major changes: the replacing of peak normalisation with z-scoring directly before feature extraction step, and resampling of the bioimpedance signals from 44100 Hz to 2048 Hz.

First, the laryngeal bioimpedance signals were subjected to additional antialiasing low-pass filtering to ensure compatibility with a reduced Nyquist frequency, corresponding to a new sampling rate of 2048 Hz. Subsequently, the bioimpedance signals were resampled at 2048 Hz, substantially reducing computational complexity, while retaining the relevant spectral content. Classification testing confirmed that this reduction in sampling rate did not

lead to a significant measurable decrease in classification accuracy, highlighting the redundancy of the higher sampling rate for this modality.

Additionally, preliminary testing suggested that peak normalisation did not adequately address sudden amplitude changes in the pathological signals, and in some cases, it failed to achieve the intended consistency across recordings. From a computational perspective, z-score normalisation offered a more robust and consistent solution for standardising the dataset. As a result, peak normalisation was excluded from the latter stage of preprocessing, and z-score normalisation was applied to all recordings immediately prior to the feature extraction phase. This approach preserved the original signal dynamics while ensuring consistency across the dataset.

4.1.4 Data Split and Categorisation

The developed database consisted of the recordings of breathing in and out through nose and mouth, coughing, sustained vowel /a/, /e/, /i/, /o/, and /u/ (each repeated at pitch C1, F1, and A1), sustained murmuring sound, as well as speech while reading one paragraph of text. Depending on the modality (audio or bioimpedance) and the content of a recording (specifically, the exact action performed by a participant such as breathing or speaking), all data underwent two stages of categorisation: the initial stage, conducted within the DAW, and the latter stage of categorising the files exported from the DAW into appropriate folders.

4.1.4.1. Initial stage

Directly after the recording process, while still in the digital audio workstation, the signals were split and categorised depending on their content, so that each file contains solely one full phonation episode. The naming of the files followed the same pattern; participant ID number, followed by “E” for laryngeal bioimpedance or “A” for audio, ending with the name

of the action performed; for instance, “b_in” for breating in, “b_out” for breathing out, “text” for speech, “tone1a” for sustained phonation of the vowel /a/ at the lowest pitch, “tone2u3” for the third attempt of sustained vowel /u/ phonation at the second lowest pitch, et cetera. For all signals obtained from the control group, the participant’s ID number is directly followed by an additional “_H”, denoting the “healthy” class.

4.1.4.2. Latter stage

Three data folders were created for each data modality: “speech” folder for the recordings of continuous speech, “sustained phonation” folder for the recordings of sustained phonation, and lastly, “others” folder for all recordings of participants’ actions unrelated directly to phonatory modes; these included all types of breathing and coughing. In the latter stage of data categorisation, all mono WAV files exported from the DAW were placed in the corresponding folder.

For the purposes of developing an accurate and reliable laryngeal pathology classification system, we decided to investigate only the phonatory data samples (only the recording of phonation); recordings of sustained vowels, including the murmuring, as well as the speech signals. Thus, the recordings of breathing and coughing were disregarded. The categorisation of the phonatory recordings into two major groups of “sustained phonation” and “continuous speech” resulted in four sub-datasets: audio sustained phonation data, audio continuous speech data, bioimpedance sustained phonation data, and bioimpedance continuous speech data.

Once fully preprocessed, the datasets consisted of the 2549 data samples in sustained phonation dataset (2549 for audio, and 2549 for EGG), and 3421 data samples in continuous speech dataset (3421 for audio, and 3421 for EGG). The following table (Table 4.3) provides the exact number of data samples in each category. All following experimentation was

pursued separately on audio data of sustained phonation, audio data of continuous speech, bioimpedance data of sustained phonation, and the bioimpedance data of continuous speech.

Table 4.3: Number of data samples in each category (after data preprocessing stage).

Category	**Audio Sustained Phonation Dataset**	**Audio Continuous Speech Dataset**	**Bioimpedance Sustained Phonation Dataset**	**Bioimpedance Continuous Speech Dataset**
Cancerous and Precancerous	759	882	759	882
Other	337	554	337	554
Neuromuscular	565	796	565	796
Laryngitis	311	575	311	575
Reinke's Oedema	292	373	292	373
Dysphonia	285	241	285	241
CONTROL (Healthy)	690	388	690	388

4.2. PRELIMINARY INVESTIGATION OF CUSTOM DATASET CLASSIFICATION

Immediately after data collection and the initial stage of preprocessing, we assessed the functionality of the custom dataset to determine whether it contained sufficient information for our intended classification tasks. This process involved a thorough statistical analysis of the gathered data and its potential to deliver sufficient information for an accurate and reliable differentiation between the investigated laryngeal pathologies. The existing laryngeal pathology classification systems rarely address the problem of speaker-dependent features and their impact on the artificially inflated overall classification performance of the developed systems. Therefore, through the preliminary investigation of the custom dataset classification, and the comparison of its results with and without the application of the

participant shuffling algorithm, we aimed at proving the negative impact of the speaker-dependent features on the overall reliability and generalisability of the developed system.

The preliminary tests were designed to evaluate the classification potential of the dataset (1), determine the impact of speaker-dependent features (2), and refine the dataset's structure to optimise classification performance (3). During the process of preliminary investigation of the data classification capabilities and the final selection of the dataset for the development of intended classification system, we prioritised two aspects:

1. The medical feasibility and necessity of laryngeal pathology classification:
    a. Which pathologies can be classified from the medical perspective?
    b. The detection of which pathologies would be most valuable for diagnostic purposes?
2. The computational feasibility of developing a reliable laryngeal pathology classification system:
    a. Does the dataset deliver sufficient information for distinguishing between the medically relevant classes?
    b. Which classes, when implemented in the designed multimodal classification system, would yield the best results from the computational perspective?

In the following section we provide the evaluation of the above considerations and the answers to the above questions. First, we discuss the preliminary investigation of the classification capabilities of the collected data. The methods used for the dataset's preliminary classification examination and the obtained results are explored. The classification model used for the preliminary investigation of the dataset's classification capabilities is briefly introduced (with its details discussed in depth in *6.2. Convolutional Neural Networks (CNN) – 6.2.2 1D-CNN – "Big" Model* section of the thesis). The preliminary investigation of the initial class setup is presented, and the obtained results are discussed,

with the particular focus on the impact of the speaker-dependent features present within data.

Subsequently, the selection of the classes for the final dataset setup is discussed. Given the medical feasibility and necessity for a particular laryngeal pathology classification system, the final subset of the collected data is chosen for the development of the intended laryngeal pathology classification system, with the final class arrangement outlined and evaluated.

In summary, in this section we outline the steps taken to evaluate the classification potential of the collected data (1), address the impact of speaker-dependent features (2), and justify the final arrangement of the dataset based on both computational and medical considerations (3).

4.2.1 Preliminary Data Classification Methods

Prior to the preliminary classification tests, the entire dataset was subjected to the initial stage of data preprocessing (*4.1.3 Data Preprocessing – 4.1.3.1. Initial Stage*); all signals of both modalities were processed in a form of 1-second-long mono WAV files with the sampling rate of 44100 Hz. Considering that the laryngeal bioimpedance signals show very little valuable spectral information above 1000 Hz as compared to audio signals (for reference, see *2.4: Investigated Laryngeal Pathologies*), the classification of electroglottographic signals with a reduced sampling rate of 2048 Hz was also performed and compared to the results of classification of the original EGG signals sampled at 44100 Hz. Reducing the sampling rate of the collected bioimpedance signals to 2048 Hz had a minimal impact on classification accuracy, but significantly reduced the computational load, validating the decision to resample the EGG signals for computational efficiency.

According to the literature and currently existing laryngeal pathology classifiers, one of the most popular and successful methods for feature extraction are spectral coefficients – most commonly, MFCCs (Arias-Londoño *et al.*, 2010; Markaki and Stylianou, 2011; Borsky *et al.*, 2017; Wang *et al.*, 2022; Miliaresi *et al.*, 2022; Islam *et al.*, 2022; Kumar *et al.*, 2023). For that reason, the preliminary classification testing was performed on the custom dataset in a form of feature sets – more specifically, the MFCCs and GTCCs, further discussed in detail in Chapter 5 of this thesis.

For the initial evaluation of the dataset's classification capabilities, a Convolutional Neural Network (CNN) with one-dimensional convolutional layers was employed – further in this study referred to as the 1D-CNN "big" model. The architecture of this 1D-CNN model is described in detail in Chapter 6 – *6.2.2 1D-CNN – "Big" Model* section of this thesis. This preliminary model served as the foundation for testing how the initial data could be used to distinguish between various laryngeal pathologies.

4.2.2 Initial Data Arrangement

Initially, the collected dataset consisted of signals gathered from seven classes of participants:

1. Control group of healthy participants,

As well as those suffering from the following laryngeal pathologies:

2. Cancerous and precancerous lesions,
3. Other benign lesions outside the vocal fold area,
4. Neuromuscular disorders,
5. Laryngitis,
6. Reinke's Oedema,
7. Functional Dysphonia.

First, all seven classes in the original setup were tested for their classification capabilities, following the program flow depicted in the following figure (Figure 4.6).

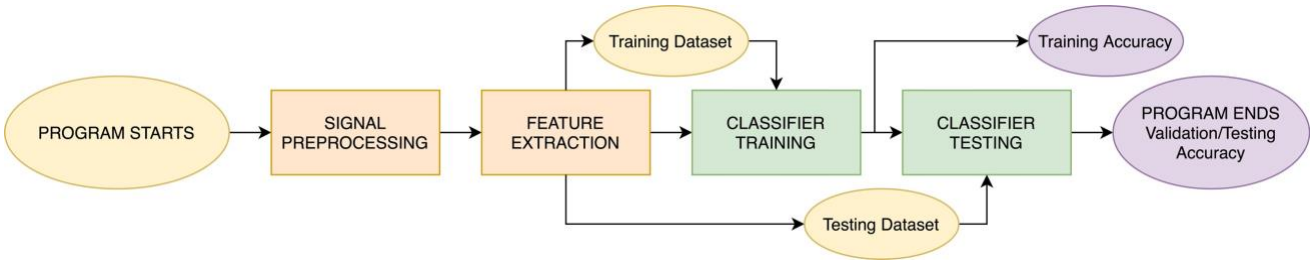

Figure 4.6: Block diagram of the processing stages of the classification model used in the first stage of preliminary classification testing of the custom dataset.

At this stage, a random shuffling approach was applied when splitting the dataset into training and validation subsets, without considering participants' identity. This meant that different recordings obtained from the same participant could appear in both training and validation sets, inadvertently allowing the classifier to learn speaker-specific characteristics rather than pathology-related differences. This was done to evidence the negative impact of the speaker-dependent features on reliability and generalisability of the designed classification system. Consequently, both continuous speech and sustained phonation data subsets delivered promising results – with the continuous speech outperforming the sustained phonation. The following table represents the average accuracy obtained for the 5-fold cross-validation preliminary classification testing with random shuffling for both audio and bioimpedance, speech and sustained phonation, tested with MFCC and GTCC as feature extraction methods:

Table 4.4: The results of the preliminary classification testing performed on the original dataset using 1D-CNN Preliminary Testing Model with random shuffling of participants. Testing was performed for EGG Sustained Phonation, EGG Speech, Audio Sustained Phonation, and Audio Speech data subsets on all six pathological classes using MFCCs and GTCCs.

	EGG sustained phonation	**EGG speech**	**AUDIO sustained phonation**	**AUDIO speech**
MFCCs	56.42% ±3.57	69.94% ±1.44	66.13% ±2.76	73.82% ±2.10

GTCCs	60.24% ±1.57	84.86% ±1.45	70.68% ±1.45	83.50% ±1.61

The following figures (Figure 4.7 and 4.8) represent confusion matrices obtained from the 1D-CNN model used for preliminary testing of multi-class classification with random shuffling; Figure 4.7 shows the best-performing model (1D-CNN fed with GTCCs derived from laryngeal bioimpedance measurements of continuous speech), and Figure 4.8 shows the worst-performing model (1D-CNN fed with MFCCs derived from laryngeal bioimpedance of sustained phonation).

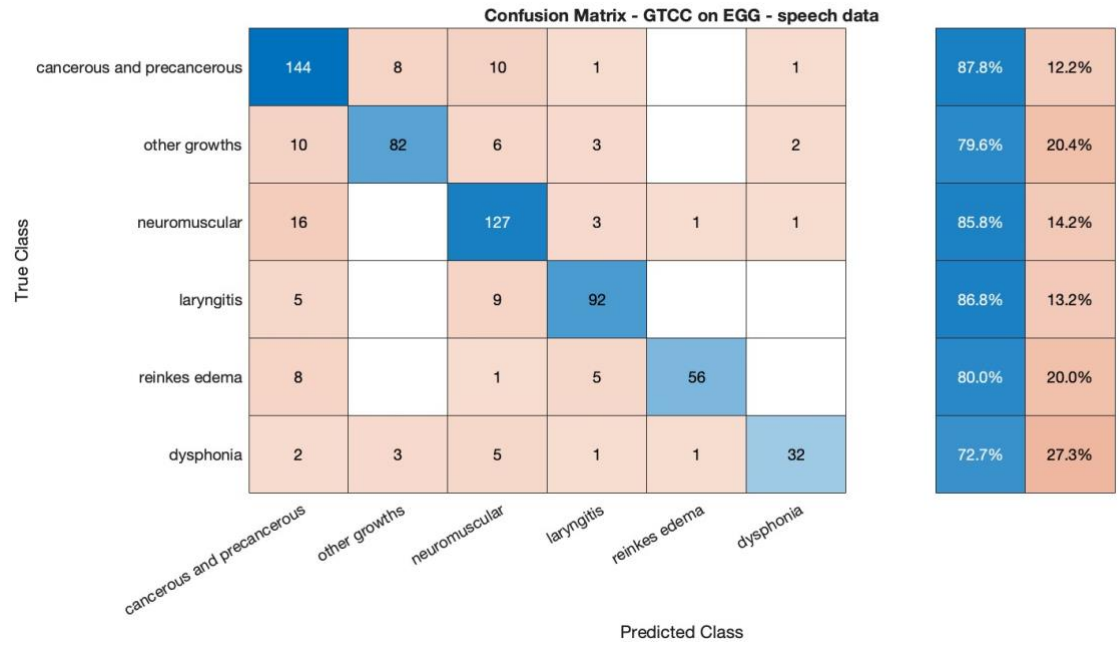

Figure 4.7: The results of the preliminary classification testing performed on the EGG Speech subset of the original dataset using GTCCs and 1D-CNN Preliminary Testing Model with random shuffling of participants.

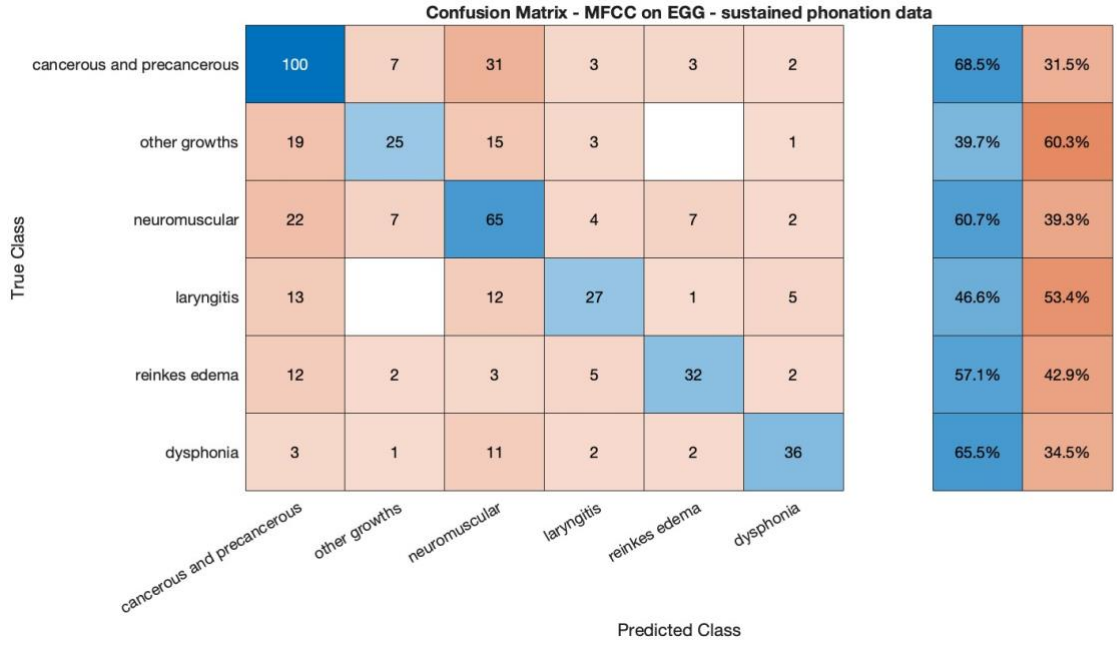

Figure 4.8: The results of the preliminary classification testing performed on the EGG Sustained Phonation subset of the original dataset using MFCCs and 1D-CNN Preliminary Testing Model with random shuffling of participants.

The initial high performance of the preliminary testing model was attributed to the presence of speaker-dependent features, which influenced the classification outcomes; the preliminary classification tests proved that speaker-specific characteristics can lead to falsely inflated classification accuracy, as initial models trained on randomly shuffled data mistakenly recognised individual participants rather than pathology-related characteristics. Nevertheless, the superior performance of continuous speech signals over sustained phonation was also recognised and noted.

4.2.3 Minimising Speaker-Dependence Bias

To address the bias introduced by speaker-dependent features, a custom shuffling algorithm was developed and implemented to ensure that samples from the same participant were exclusively placed in either training or validation sets (with the ratio of 80%-20%), but never in both simultaneously. The new classification model followed the processes depicted in the following flow chart (Figure 4.9).

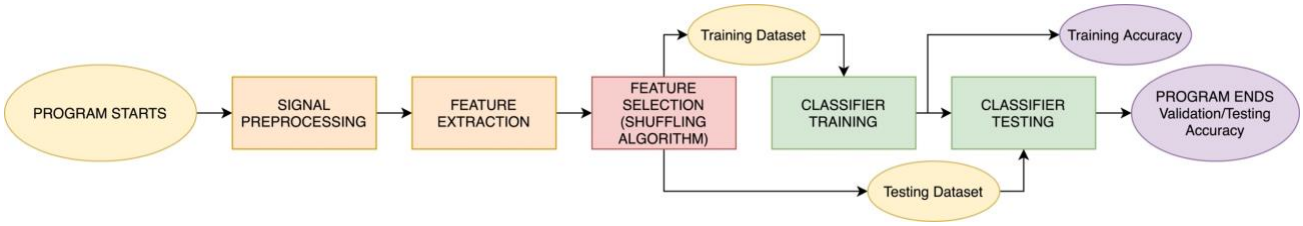

Figure 4.9: Block diagram of the processing stages of the classification model used in the second stage of preliminary classification testing, as well as in the final classification system proposed in this research.

As expected, while minimising the data bias and the impact of the speaker-dependent features on the classifier' performance, the use of the participant shuffling algorithm significantly reduced the classification accuracy. This further proved that the initial high performance of the model was due to overfitting to speaker identity rather than the pathology-related classification.

Under the more rigorous data shuffling, that prevented the speaker-dependent features from influencing the classification performance, certain classes, such as laryngitis and benign lesions outside the glottal area, were classified at a random rate. Furthermore, misclassifications were observed; the first major misclassification was dysphonia cases assigned to the neuromuscular disorders – this can be attributed to dysphonia being a symptom of neuromuscular disease. Another example of misclassification was laryngitis – this disorder constitutes an inflammation of the laryngeal tissues. Laryngitis, as a common disorder co-occurring with various laryngeal growths (Soni *et al.*, 2016; Jetté, 2016), constituted another major misclassification.

These findings highlighted the lack of consistent aetiology in some of the original classes, and by extension, the lack of consistent features for their accurate classification. Furthermore, from the medical perspective, some of the original pathology classes could not be assigned to a single defined class without simultaneously belonging to another. It became evident that the seven-class classification approach was neither practical nor clinically

meaningful. This led us to believe a more robust approach to the pathological class arrangement was required.

4.2.4 Medical Feasibility Analysis

Based on both the computational results and medical considerations, several of the initial classes were re-evaluated:

4.2.4.1. Dysphonia

Although present in the dataset, dysphonia was ultimately excluded from the final dataset setup, since in most medical cases, dysphonia is considered a symptom of an underlying disorder (Roy, 2003), not a disease entity or a standalone diagnosis (Przysiezny and Przysiezny, 2015). Moreover, dysphonia can be linked to non-organic factors such as psychological trauma or emotional stress, which do not involve direct physical changes in the larynx – studies report that emotional stress and anxiety are severely more common in subjects suffering with dysphonia than in general population (Ferrán *et al.*, 2024; Martins *et al.*, 2014). The focus of this research is strictly on pathologies resulting from physical or physiological impairments as standalone diagnoses – therefore, dysphonia was finally excluded from the final study.

4.2.4.2. Laryngitis and Other Benign Lesions

Similarly, laryngitis and other benign lesions outside the vocal fold area were excluded from the final dataset setup due to difficulty in categorising them under a standalone diagnosis. Medically, laryngitis is frequently a symptom of broader inflammatory processes, often caused by a different underlying pathology (Jetté, 2016). Because inflammation frequently co-occurs with other lesions, laryngitis is often present alongside other laryngeal pathologies; for instance, chronic laryngeal inflammation is commonly observed in patients

who also have benign vocal fold lesions (Jetté, 2016). In practice, this means a patient with another vocal cord problem often has laryngitis simultaneously, as the lesion can provoke local inflammation and vice versa. Therefore, its presentation can be nonspecific, and its classification as a sole diagnosis, unreliable (Jetté, 2016).

According to literature, benign lesions located outside the true vocal folds have minimal impact on vocal fold vibratory behaviour and voice acoustics (Lee *et al.*, 2019; Shrivas *et al.*, 2022), therefore, voice-based pathology classification models will struggle to detect them reliably. Consultation with a medical professional confirmed that these classes could be misleading and did not offer the clear diagnostic value needed for the intended classification system. Reinke's Oedema, while initially considered, was also re-evaluated under these criteria.

4.2.4.3. Priority for Cancer Detection

There is a significant clinical need for the early detection of cancerous and precancerous lesions (DuBois, 2021). This medical urgency, combined with the feasibility of extracting reliable features from the data, led to the decision to prioritise cancer detection alongside neuromuscular disorders.

It is important to acknowledge that cancerous and precancerous lesions were treated as a single category in this research. While clinically distinct, their separation requires careful medical supervision, invasive diagnostic methods, and – to some extent – it remains subjective. Within the scope of non-invasive, digital signal-based approaches such as those investigated here, the prevailing view is that these methods serve primarily as pre-screening tools, identifying individuals who may require further, invasive investigation. In this context, a combined cancerous-precancerous category is sufficient to meet the primary aim: distinguishing patients who may benefit from urgent medical referral from those who are

unlikely to require it. Future work, supported by larger datasets and extended clinical input, could aim to separate cancerous and precancerous cases.

4.2.5 Final Dataset Arrangement

Considering both the computational experiments and the medical feasibility analysis, the final dataset was restructured to focus on three clinically significant pathology groups:

- Cancerous and precancerous lesions – conditions with malignant potential, requiring early detection for timely intervention.
- Neuromuscular disorders – primarily vocal fold paralysis, representing disorders with distinct biomechanical characteristics.
- Healthy controls – participants without any laryngeal pathology, serving as a control group.

This focused subset ensured that the subsequent data analysis and model development are both clinically relevant and computationally robust, addressing the critical need for early cancer detection while providing a clear framework for distinguishing neuromuscular pathology from normal conditions. This reorganisation was based on the principle that the primary objective of the classification system was to prioritise the detection of malignancies, rather than attempting to distinguish between overlapping non-cancerous conditions. The transition to a three-class classification approach yielded significantly improved classification performance (further explored in Chapters 8 and 9), as it eliminated redundant and overlapping categories that previously hindered model learning.

However, for the development of the intended laryngeal pathology detection system (the binary classification between healthy and pathological cases), all pathological classes were considered. This decision was made in line with the medical feasibility analysis, and the purposes of the designed system.

4.3. EXPLORATORY DATA ANALYSIS

In this section, we analysed the recorded audio and laryngeal bioimpedance signals using statistical and spectral methods to understand the distribution of data, identify the trends and the presence of potential clustering tendencies, and evaluate the class separability before building the intended classification system. For that, the sets of statistical and time- and frequency-domain measures were derived from the data and analysed for the presence of features relevant for the identification of specific laryngeal pathologies and distinguishing between pathological and healthy signals.

One of the key aspects of this analysis was to determine whether the low-complexity statistical measures are sufficient for a reliable pathology classification or whether the more advanced computational approaches – particularly deep learning-based classification – are necessary. By exploring class separability using relatively low-cost statistical techniques, this section provides insights into the limitations of conventional classification methods and justifies the need for more robust learning-based approaches.

The exploratory data analysis (EDA) involved both statistical analysis and visualisation techniques to examine the collected custom dataset. First, we analysed the descriptive statistics of the extracted features to observe trends across different data classes. Subsequently, PCA was performed to explore class separability in both the time-domain and the frequency-domain. PCA helped determining how well the investigated classes cluster in lower-dimensional space, offering insights into the feature significance (Abdi and Williams, 2010). Finally, based on the two first principal components, the Hopkins statistics were calculated for each class of each data modality to assess its clustering tendencies (Hopkins and Skellam, 1954).

Following the conclusion drawn in *4.2. Preliminary Investigation of Custom Dataset Classification* section, it is evident that the speaker-specific characteristics may lead to

falsely inflated accuracy of the pathology classification. Therefore, the following data analysis was performed on the classes resulting from the final rearrangement of the dataset's setup intended to fulfil the primary goal of the envisaged classification model – the laryngeal pathology classification between cancerous and precancerous lesions, neuromuscular disorders, and healthy cases, with the particular focus on the detection of malignant cases for the early intervention.

This subchapter is structured as follows:

- *4.3.1 Overview of Data Analysis Methods and Results*: The overview of parameters used to evaluate the data, followed by the overview of most prominent tendencies discovered through the data analysis.
- *4.3.2 Investigation of Class Separability using Global Statistical Features and PCA*: Examining class separation using primarily time-domain and statistical methods.
- *4.3.3 Investigation of Class Separability using Time-Frequency Parameters and PCA*: Assessing separability in the spectral domain.
- *4.3.4 Investigation of Class Separability using Combined Parameters and PCA*: Evaluation of data separability using Global Statistical Features and the Time-Frequency Parameters combined.
- *4.3.5 Combined Multimodal Analysis of Class Separability using PCA*: Evaluation of data separability using both Global Statistical and Time-Frequency Parameters derived from the combined investigated data modalities – audio and laryngeal bioimpedance concatenated together.
- *4.3.6 Summary of Findings and Implications for Model Development*: The summary of the data analysis results and conclusions, containing brief

descriptions of most crucial data tendencies and the resulting approach to data classification.

4.3.1 Overview of Data Analysis Methods and Results

This subsection discusses in detail the methods that were applied for the data analysis. The analysis was performed using MATLAB computing environment (the code written for the data analysis can be found in the algorithms folder created for the purposes of this research that is available upon request).

First, the signals of both data modalities from all classes were loaded into MATLAB in a form of audio datastores for the ease of data processing. All parameters used for the extraction of features were subsequently declared; sampling rate (*Fs = 44100*), frame size (*frameSize = 512*), rectangular window (*rect_win = rectwin(frameSize)*), overlap (*overlap = round(0.5*frameSize)*).

Directly before obtaining the signal characteristics, the z-score normalisation was performed to standardise the dataset. Notably, the mean and standard deviation (SD) of the z-scored data were also calculated and included as additional features in the performed PCA. This was done to serve as a safeguard, as the mean and SD values of the z-scored data are expected to be constant across all samples. As such, these features should contribute minimally to the derivation of the principal components, allowing the PCA to focus on the more informative variations in the data.

The extracted measures were grouped into two categories: first of Global Statistical Features, investigating signals in their entirety with no windowing; second being Time-Frequency Parameters, evaluating the signals in segments using windowing to assess their spectral properties.

4.3.1.1. Global Statistical Features

In this method, the “features” (statistical parameters) were derived from the entire signal without dividing it into segments. Features like mean, SD, skewness, and kurtosis were calculated across the entire signal duration. The mean and SD were included to serve as a safeguard, given all data was z-scored prior to the PCA. Consequently, these parameters were expected to contribute minimally to the derivation of the principal components.

The “Global Statistical Features” approach was enriched with the following measures, also calculated over the entire signal duration: signal to noise ratio (SNR), total harmonic distortion (THD), harmonic ratio (HR), and the autocorrelation features. To achieve a more comprehensive analysis, we augmented our Global Statistical Feature set with two additional features: harmonic-to-noise ratio (HNR) and magnitude spectrogram calculated using the short-time Fourier Transform (STFT). HNR was included to complement the existing SNR and harmonic ratio measures; to avoid replicating SNR values, the HNR was computed using the Boersma approach (1993) of autocorrelation of a windowed signal, by taking the mean of values calculated across the signal. STFT magnitude statistics, expressed as their mean and standard deviation, summarise the overall energy distribution of the signal. They were incorporated to capture a more holistic view of the signal’s dynamic variations.

The following table (Table 4.5) lists all the parameters derived using the Global Statistical Feature approach, providing their short definition and mathematical formula. The parameters are calculated for:

- x representing the signal,
- n representing the sample number of a signal, where x_n stands for the n-th sample of the analysed signal x,
- N representing the length of a signal,

- f_0 corresponding to fundamental frequency,
- P_x corresponding to the power of the signal x, with P_{signal} corresponding to the power of a desired signal (fundamental frequency and first five harmonics), P_{noise} corresponding to the power of noise, P_{har} corresponding to the power of harmonic components (other than the fundamental frequency), and P_{f_0} referring to the power of the fundamental frequency, where

$$P_x = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |x[n]|^2 \tag{4.1}$$

- τ corresponding to the time-lag used in the autocorrelation function,
- $r_x(\tau)$ representing the autocorrelation function of the signal x,
- $r'_x(\tau)$ representing the autocorrelation function of the signal x normalised by its global maximum $r_x(0)$, like so:

$$r'_x(\tau) = \frac{r_x(\tau)}{r_x(0)} \tag{4.2}$$

- $r'(\tau_{max})$ representing the maximum of the normalised autocorrelation function calculated for signal x for $\tau > 0$, which corresponds to the power of harmonic component,
- k is the number of the frequency bin,
- f_k representing the frequency in Hz corresponding to k-th bin,
- m standing for time window,
- $X(m, f_k)$ representing corresponding spectral magnitudes of k-th bin and m time window, calculated using STFT:

$$X(m, f_k) = \sum_{n=-\infty}^{\infty} x[n]w[n - mH]e^{-j2\pi f_k n} \tag{4.3}$$

where $w[n]$ is the window function with H hop size.

Table 4.5: Mathematical and conceptual definition of all measures derived in the Global Statistical Features approach to statistical analysis.

Measure's Name	**Measure's Description**	**Measure's Formula**
Mean	The average value across all values in the dataset – in this case, it is the average amplitude level across the entire signal. Due to z-scoring, all mean values calculated for the signals in the dataset are equal to 0.	$mean(x) = \mu = \frac{1}{N}\sum_{n=1}^{N} x_n$ (4.4)
Standard Deviation (SD)	Captures how spread out the signal's values are around its mean. In this case, the values of the signal's amplitude. Due to z-scoring, all SD values calculated for the signals in the dataset are equal to 1.	$SD(x) = \sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2}$ (4.5)
Skewness	Represents the degree of asymmetry of the signal's distribution (Groeneveld and Meeden, 1984). The positive skewness stands for the right-skewed distribution, while negative – left-skewed.	$skew(x) = \gamma_1 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n - \mu}{\sigma}\right)^3$ (4.6)
Kurtosis	Represents the degree of "peakedness" of the signal's distribution – measures how heavily the tails of the distribution deviate from the normal (Gaussian) distribution (Groeneveld and Meeden, 1984). High kurtosis suggests heavy tails and a distinct peak.	$kurt(x) = \kappa = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n - \mu}{\sigma}\right)^4$ (4.7)
Signal to Noise Ratio (SNR)	Ratio in dB of the power of the desired signal and the power of the background noise (Oppenheim and Schafer, 1989). For the purposes of this study, the fundamental frequency and the first five harmonics were treated as the desired signal, while the remaining frequency components were treated as the noise. The SNR is determined using a modified periodogram of the same length as the input and a modified periodogram with a Kaiser window of 38 sample sidelobe attenuation of the Fourier transform.	$SNR = 10 \cdot \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$ (4.8)
Total Harmonic Distortion (THD)	Measures harmonic distortion in dB present in a signal, calculated as ratio of the sum of powers of all harmonic components other than fundamental, and the power of the fundamental (Blagouchine and Moreau, 2011). Higher THD indicates more distortion. For the purposes of this study, the THD was determined from the fundamental frequency and the first five harmonics.	$THD = 10 \cdot \log_{10}\left(\frac{P_{har}}{P_{f_0}}\right)$ (4.9)
Harmonic Ratio (HR)	Measures the proportion of harmonic components in the power spectrum (Kim ***et al.***, 2006) as the ratio of harmonic energy to the total energy of the signal.	$HR = \frac{P_{f_0} + P_{har}}{P_x}$ (4.10)

Autocorrelation Features	Examines how the signal correlates with time-shifted version of itself, where peaks in the positive-lag region reveal periodic and quasi-periodic patterns (Boersma, 1993). Represented as mean and SD of the peaks in the function at positive lags. These measures capture the strength and consistency of signal's periodicity.	$NormAutoCorr = \frac{\sum_{n=1}^{N-\tau} x_n x_{n+\tau}}{\sqrt{\sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} x_{n+\tau}^2}}$ (4.11)
Harmonic to Noise Ratio (HNR)	Ratio in dB of the periodic (harmonic) energy to nonperiodic (noise) energy in the signal (Fernandes *et al.*, 2018). For the purposes of this study, the HNR was calculated following the Boersma's approach (Boersma, 1993), where for each short-time frame, we window the signal and compute the autocorrelation via FFT→ |magnitude|² → IFFT, normalise the autocorrelation function, and divide its peak $r'(\tau_{\max})$ by the power of remaining, non-harmonic component.	$HNR = 10 \cdot \log_{10}\left(\frac{r'(\tau_{\max})}{1 - r'(\tau_{\max})}\right)$ (4.12)
STFT Magnitude Statistics	Returns Short-time Fourier Transform magnitude values calculated across all time-frequency bins, representing the overall energy distribution of the signal across the time-frequency domain. Represented as mean and SD of all values calculated across the signal.	$MagnitudeSpec = \|X(m, f_k)\|$ (4.13)

4.3.1.2. Time-Frequency Parameters

This approach involved first segmenting the signal into smaller windows and then computing time-frequency domain parameters – in MATLAB referred to as "spectral descriptors" (MathWorks, 2024) – including: spectral centroid, spectral entropy, spectral roll-off point, spectral spread, spectral skewness, spectral kurtosis, spectral flatness, spectral crest, spectral flux, spectral slope, and spectral decrease. Subsequently, the mean and standard deviation were calculated across these parameters.

The following table (Table 4.6) provides a short description and a mathematical formula / criterion for all time-frequency parameters calculated for the purposes of data analysis. All the following descriptors refer to the distribution of the energy across the analysed signals. The parameters are defined for:

- k is the number of the frequency bin,
- f_k representing the frequency in Hz corresponding to k-th bin,
- $X(f_k)$ representing corresponding spectral magnitudes of k-th bin,
- p_k corresponding to normalised spectral magnitude,
- m standing for time window,
- N representing the total number of samples in the signal or frequency bins in the spectral representation,
- $\bar{f}$ corresponding to mean frequency,
- $\bar{X}$ corresponding to mean spectral magnitude.

Table 4.6: Mathematical and conceptual definition of all measures derived in the Time-Frequency Parameters approach to statistical analysis.

Measure's Name	**Measure's Description**	**Measure's Formula**
Spectral Centroid	The frequency-weighted sum normalised by the unweighted sum (Peeters, 2004). Often referred to as the "centre of gravity" of the spectrum, considered a measure of brightness of the sound – higher centroid indicates more energy of the signal is distributed across high frequencies.	$\mu_1 = \frac{\sum_{k=1}^{N} f_k X(f_k)}{\sum_{k=1}^{N} X(f_k)}$ (4.14)
Spectral Entropy	Indicative of the peakiness of the spectrum, quantifies the spectrum disorder (Misra *et al.*, 2004). Higher entropy indicates a more uniform energy distribution, akin to noise, lower entropy corresponds to more periodic signal, alike speech or sustained phonation.	$entropy = -\sum_{k=1}^{N} p_k \log_2(p_k)$ where $p_k = \frac{X(f_k)}{\sum_{k=1}^{N} X(f_k)}$ (4.15)
Spectral Roll-off Point	Measures the frequency bandwidth under which a given percentage (normally set to 85-95%) of the total energy resides (Scheirer and Slaney, 1997).	$rolloff = i, for \sum_{\{f \le f_r\}}^{i} X(f) \ge \alpha \sum_{all\ f} X(f)$ (4.16)
Spectral Spread	Represents the SD around spectral centroid (Peeters, 2004) – reflects how broad or narrow the spectrum is around its centroid. Large spectral spread indicates the energy of the signals is more widely dispersed across frequencies.	$\mu_2 = \sqrt{\frac{\sum_{k=1}^{N} (f_k - \mu_1)^2 X(f_k)}{\sum_{k=1}^{N} X(f_k)}}$ (4.17)

Spectral Skewness	Computed from the third order moment (Peeters, 2004), spectral skewness identifies whether most spectral energy lies below or above the centroid. A positive skew indicates a heavier tail in the upper frequency range.	$\mu_3 = \frac{\sum_{k=1}^{N}(f_k - \mu_1)^3 X(f_k)}{(\sum_{k=1}^{N} X(f_k))\mu_2^3}$ (4.18)
Spectral Kurtosis	Computed from the fourth order moment (Peeters, 2004), spectral kurtosis measures flatness or non-Gaussianity of the spectrum around its centroid – identified whether the spectrum has heavy tails or sharp peaks. High value indicates a more peaked spectrum.	$\mu_4 = \frac{\sum_{k=1}^{N}(f_k - \mu_1)^4 X(f_k)}{(\sum_{k=1}^{N} X(f_k))\mu_2^4}$ (4.19)
Spectral Flatness	Measures ratio of the geometric mean of the spectrum to its arithmetic mean (Johnston, 1988) – another indication of the spectrum peakiness. A higher spectral flatness corresponds to the presence of noise, lower indicates more tonality.	$flatness = \frac{\exp\left(\sum_{k=1}^{N} f_k \ln\left(X(f_k)\right)\right)}{\frac{1}{N}\sum_{k=1}^{N} X(f_k)}$ (4.20)
Spectral Crest	Calculates ratio between the maximum of the spectrum to its arithmetic mean (Peeters, 2004), as another measure of peakiness. Higher crest indicates more tonality, lower – more noise.	$crest = \frac{\max_k X(f_k)}{\frac{1}{N}\sum_{k=1}^{N} X(f_k)}$ (4.21)
Spectral Flux	Measure of the variability of the spectrum over time (Scheirer and Slaney, 1997) – quantifies how much spectrum changes from one window to another. A high flux value indicates abrupt time-frequency variations.	$flux(t) = \sum_{k=1}^{N}\left(X(m, f_k) - X(m-1, f_k)\right)^2$ (4.22)
Spectral Slope	Evaluates the general incline of decline of spectral amplitude with increasing frequency (Lerch, 2012). A steep negative slope indicates the energy quickly diminishes in higher frequencies. Spectral slope is believed to be directly related to the resonant frequencies of the vocal folds and has been applied in modelling speaker stress (Hansen and Patil, 2007).	$slope = \frac{\sum_{k=1}^{N}(f_k - \bar{f}) \cdot (X(f_k) - \bar{X})}{\sum_{k=1}^{N}\left(f_k - \bar{f}\right)^2}$ (4.23)
Spectral Decrease	Measures how amplitudes decrease as frequency increases, emphasising the slopes of the lower frequencies, and providing another angle at spectral tilt (Peeters, 2004). A strong spectral decrease suggests the energy drops rapidly after the fundamental or lower overtones.	$decrease = \frac{\sum_{k=2}^{N} \frac{X(f_k)}{k-1}}{\sum_{k=1}^{N} X(f_k)}$ (4.24)

4.3.1.3. Methods of Cluster Separability Assessment

Following the analysis of both groups of features derived for all classes of investigated pathologies, PCA was performed and analysed both statistically and visually. Subsequently, Hopkins statistics were calculated for the first two principal components (PC) of each class

to assess their clustering tendencies. Furthermore, based on the first three PCs of each class, we calculated Euclidean distances between class centroids to present the separability of the clusters according to the ground truth labels.

The PCA is a statistical technique used to reduce data's dimensionality while retaining most important information within, preserving its variance (Abdi and Williams, 2010). By transforming the original feature space into a new set of orthogonal components – the *principal components* – PCA captures the directions of maximal variance, with the first principal component accounting for the largest variance, the second for the next largest, and so on (Abdi and Williams, 2010). It is therefore commonly used to find and represent the underlying structure in high-dimensional data. By projecting the data onto the computed principal components, PCA yields a lower-dimensional representation that preserves as much variance in the data as possible.

Introduced by Hopkins and Skellam (1954), the Hopkins statistic is a measure of "clusterability" (data's clustering tendency or its non-random structure) used for datasets in cluster analysis. It compares the distribution of the distance between real data points and their nearest neighbours to the distribution of the distance of randomly generated points (in the same space) and their nearest neighbours. Given it assumes the null hypothesis where the data points are uniformly randomly distributed, the value dropping below 0.5 suggests normal distribution, and the value approaching 1 indicates a strong clustering tendency (Hopkins and Skellam, 1954).

Hopkins statistics are calculated by comparing the sum of distances between the n real data points and their nearest neighbours with the sums of distances for the n random points and their nearest neighbours. If a data subset of n real points is randomly selected (where X is the set of selected data points x_i), and the n truly random points from the original data

space's bounding region are generated (where Y is the set of random points y_i), then the Hopkins statistic H is equal to:

$$H = \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} u_i + \sum_{i=1}^{n} w_i} \tag{4.25}$$

Where u_i is the minimum distance of y_i to its nearest neighbour from X, and w_i is the minimum distance of x_i to its nearest neighbour from X (Hopkins and Skellam, 1954).

For the purposes of this data analysis, the value of n is chosen as a fixed number of 50, since it is generally considered as large enough to capture the data's tendency toward clustering without incurring excessive computation.

To provide further insight into the class separability, we calculated Euclidean distances between class centroids, thus, the length of a straight-line distance between the centre point of the clusters calculated as the mean of all data points within that cluster.

4.3.1.4. Preliminary Results of Data Analysis

A summary of the parameters derived for the purposes of the data analysis is presented in Table 12.1 in the *Appendices* section, that shows the values for all Global Statistical Features and Time-Frequency Parameters extracted across the two investigated pathologies and the healthy group. The following most crucial trends were observed in the dataset:

- Overall, the data analysis revealed notable statistical differences in parameters between the pathological and healthy groups. However, the distinctions between the pathological conditions were much less pronounced.
- The spectral measures, including spectral entropy and spectral spread, showed most notable differences between healthy and pathological cases, with pathologies exhibiting much broader spectral content and higher entropy values.

Noticeably, the highest values of spectral descriptors including entropy, roll-off point, spread, and flatness, were achieved for malignant cases in audio, and neuromuscular cases for laryngeal bioimpedance.

- For laryngeal bioimpedance, the spectral measures of spectral centroid, entropy, flatness, roll-off point, and spread were noticeably higher for neuromuscular disorders than any other class, with spectral spread and roll-off point reaching over double of that calculated for the healthy cases (the mean of spectral roll-off for healthy = 742 Hz; for neuromuscular = 2354 Hz; the mean of spectral spread for healthy = 576 Hz; for neuromuscular = 1286 Hz). The values of SNR and HNR were also the lowest for the neuromuscular class, indicating reduced clarity of the underlying tonal characteristics and suggesting those signals are heavily influenced by noise. The SNR and HNR were the highest for the bioimpedance healthy signals.
- Both modalities showed increased HNR, autocorrelation, and harmonic ratio (with the lowest SD), the highest spectral crest, and the lowest spectral entropy and spectral flatness for healthy signals, indicating clearer, more harmonic-rich signals with well-defined periodicity and tonality.

4.3.2 Investigation of Class Separability using Global Statistical Features and PCA

To assess class separability using the Global Statistical Features, the PCA was applied to the z-score normalised feature set derived separately from audio and the laryngeal bioimpedance. The first two principal components (PC1 and PC2) were examined to visualise the clustering tendencies among the investigated classes of both data modalities (Figure 4.12 and 4.13). Figure 4.10 presents a Pareto bar chart illustrating the initial principal

components accounting for over 95% of the total variance in both the audio and bioimpedance modalities.

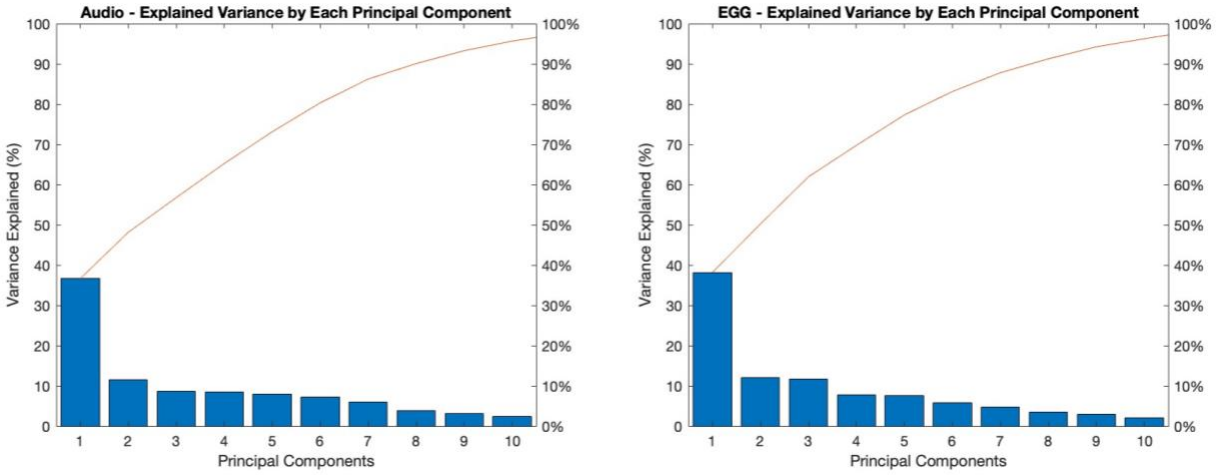

Figure 4.10: Pareto bar chart of first principal components explaining 95% of variance in Global Statistical Features derived from both data modalities – audio and laryngeal bioimpedance (EGG).

According to the evaluation of class separability using Global Statistical Features approach, the parameters that were most impactful for the calculation of the first principal component were harmonic to noise ratio, autocorrelation features (mean and SD), signal to noise ratio, and mean values of STFT magnitude features. Naturally, following the z-scoring, the values of mean and SD calculated for the signals were least valuable for the explanation of data variance. Figure 4.11 represents the contributions of the global signal characteristics into the first principal component.

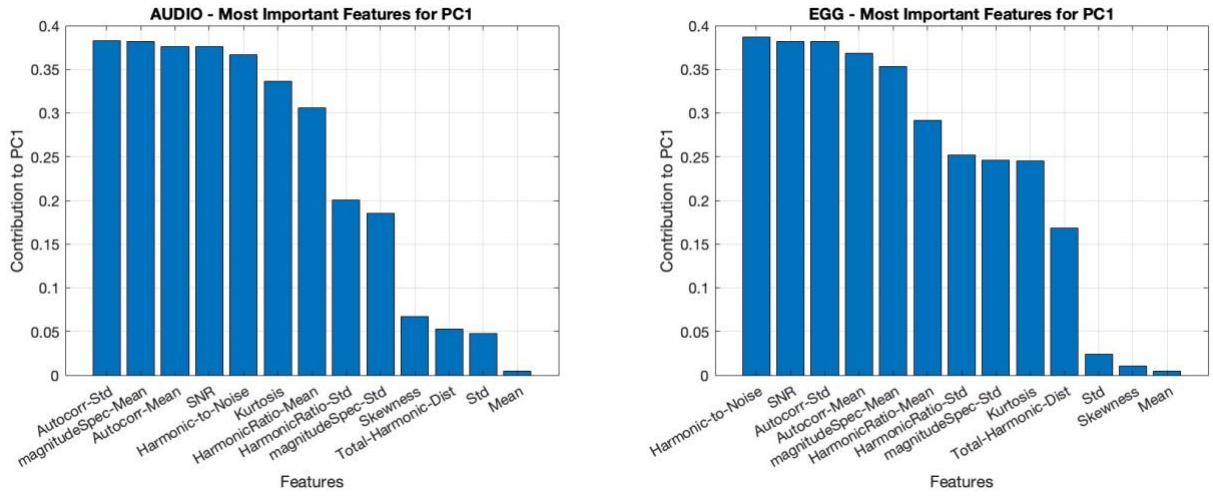

Figure 4.11: Contribution of the parameters from the Global Statistical Features set to the first principal component for both modalities. "Std" stands for standard deviation (SD), "Autocorr-Mean" and "Autocorr-Std" stand for mean and SD values of autocorrelation features, "SNR" stands for signal to noise ratio, "magnitudeSpec-Mean" and "magnitudeSpec-Std" signifies respectively mean and SD of STFT magnitude features, and "Total-Harmonic-Dist" stands for the total harmonic distortion.

4.3.2.1. Audio-based PCA

The PCA results for Global Statistical Features derived from the audio signals showed moderate class separability, with the healthy group clustering distinctly from malignant and neuromuscular conditions. However, a significant overlap between the two investigated pathologies was observed, indicating a possible lack of sufficient differences in the derived Global Statistical Features to distinguish between the two classes.

The following figure (Figure 4.12) shows the distribution of the audio signals projected onto the PCA space using first two principal components, colour-coded by class. Convex hulls are drawn around each class to illustrate their boundaries and how well the data points are separated in this reduced-dimensional space.

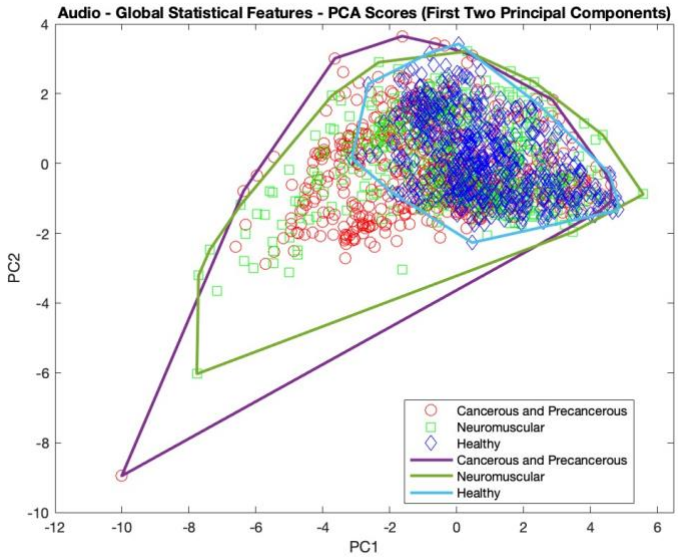

Figure 4.12: PCA score plot of Audio-derived Global Statistical Features: Audio dataset of three investigated classes projected within PCA space using first two PCs derived from Global Statistical Features. The boarders represent the convex hulls derived for each class in the two-dimensional space.

Hopkins statistic calculated for the first two PCs derived from the Global Statistical Feature set of audio signals of the three investigated classes confirmed the data exhibits moderate to high clustering behaviour – the calculated Hopkins H values were:

- For class of cancerous and precancerous lesions: 0.9500,
- For class of neuromuscular disorders: 0.8910,
- For healthy class: 0.8240.

The Pairwise Euclidean Distances between three classes' centroids calculated for audio signals in the PCA space were as follows:

- Cancerous and precancerous vs Neuromuscular: 0.6201,
- Cancerous and precancerous vs Healthy: 1.5583,
- Neuromuscular vs Healthy: 0.9426.

4.3.2.2. Bioimpedance-based PCA

The PCA performed on the Global Statistical Features derived from laryngeal bioimpedance signals revealed even better separability between healthy and pathological cases, particularly along PC1. This suggests that EGG signals may capture distinctive acoustic characteristics of the vocal fold vibration patterns that are especially relevant for the detection of a pathology – possibly with an advantage over the audio signals. Although with clustering tendencies showing slightly distinct directions, the malignant and neuromuscular classes show a considerable overlap.

Figure 4.13 displays the bioimpedance signals of all three classes projected onto the first two principal components, with each class distinguished by colour. Convex hulls enclose the points for each class, illustrating their boundaries and highlighting the degree of separation in the PCA space.

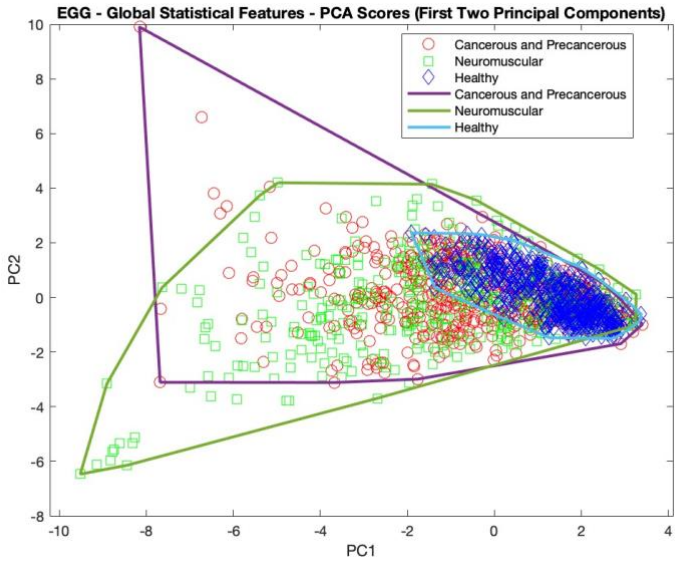

Figure 4.13: PCA score plot of EGG-derived Global Statistical Features: Bioimpedance dataset of three investigated classes projected within PCA space using first two PCs derived from Global Statistical Features. The boarders represent the convex hulls derived for each class in the two-dimensional space.

Hopkins statistic computed for the first two PCs derived from the Global Statistical Feature set of laryngeal bioimpedance signals of the three investigated classes indicates

that the data exhibits well-defined clustering behaviour. The corresponding values of H were:

- For class of cancerous and precancerous: 0.9592,
- For class of neuromuscular: 0.8319,
- For healthy signals: 0.8331.

The Pairwise Euclidean Distances between the centroids calculated for the three classes of bioimpedance signals in the PCA space were as follows:

- Cancerous and precancerous vs Neuromuscular: 0.7361,
- Cancerous and precancerous vs Healthy: 1.5811,
- Neuromuscular vs Healthy: 2.2548.

4.3.3 Investigation of Class Separability using Time-Frequency Parameters and PCA

Following the approach employed in the Global Statistical Features analysis, the data analysis using Time-Frequency Parameters was also evaluated by applying PCA to z-score normalised feature sets derived separately from audio signals and laryngeal bioimpedance. Using PC1 and PC2 the clustering tendencies of the investigated classes and derived parameters were visualised in both data modalities (Figure 4.16 and 4.17). Figure 4.14 depicts a Pareto bar chart highlighting the first principal components that together account for over 95% of the total variance in both the audio and bioimpedance modalities.

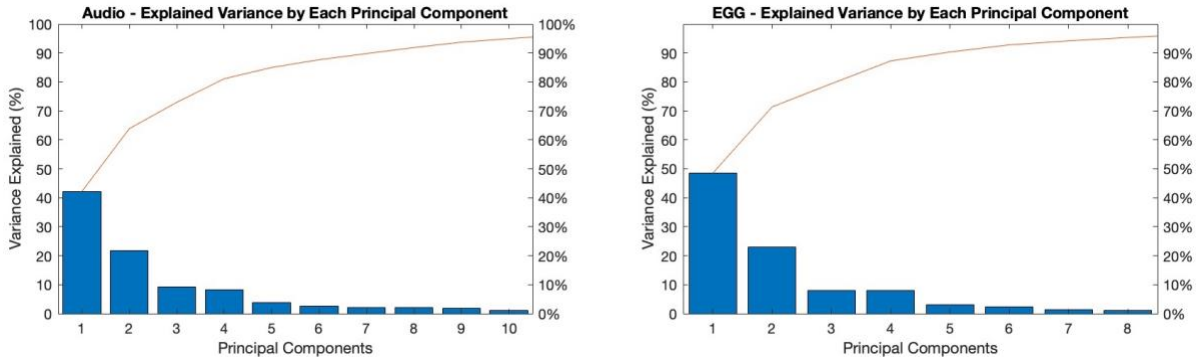

Figure 4.14: Pareto bar chart of first principal components explaining 95% of variance in Time-Frequency Parameters derived from both data modalities – audio and laryngeal bioimpedance (EGG).

From this class separability assessment, the parameters most influential in forming the first principal component were spectral flatness (both its mean and the SD), spectral spread, and spectral entropy, with spectral roll-off point and the SD calculated for spectral centroid following as the fifth or sixth parameter. Figure 4.15 illustrates the contributions of each Time-Frequency Parameter to the first principal component.

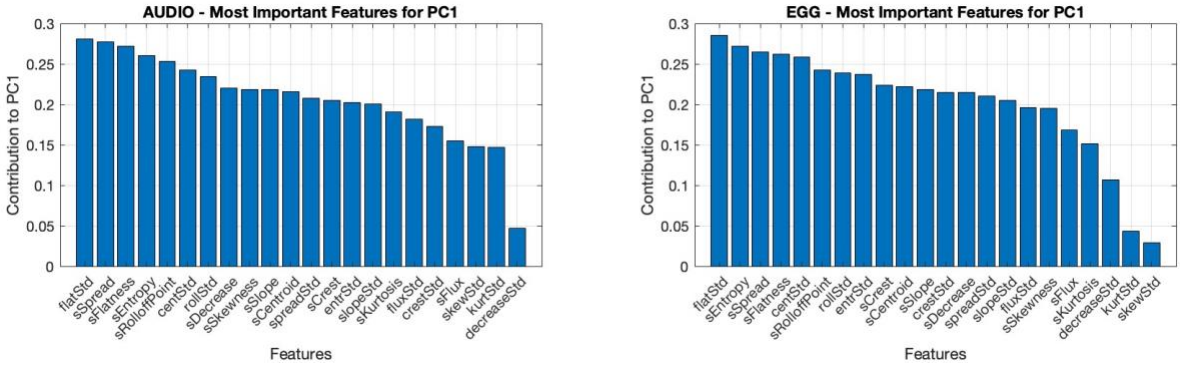

Figure 4.15: Contribution of the parameters from the Time-Frequency Parameters set to the first principal component for both modalities. The "s" preceding the name of the parameters stands for "spectral", while "Std" succeeding the parameters' name stands for standard deviation, where "flat" is spectral flatness, "cent" is spectral centroid, "roll" is spectral roll-off point, "spread" is spectral spread, "entr" is spectral entropy, "slope" is spectral slope, "flux" is spectral flux, "crest" is spectral crest, "skew" is spectral skewness, "kurt" is spectral kurtosis, and "decrease" is spectral decrease.

4.3.3.1. Audio-based PCA

The spectral content of audio signals projected onto the PCA space improved slightly the class separability compared to the Global Statistical Features, with spectral flatness and spectral spread contributing significantly towards PC1. Healthy and pathological signals

displayed visibly distinct clustering tendencies, however, cancerous-precancerous lesions and neuromuscular disorders remained overlapped. This indicates a slight statistical difference in the derived Time-Frequency Parameters between the two investigated pathologies which is, however, not sufficient for the accurate classification between the two cases based solely on their Time-Frequency Parameters and statistical classification methods.

Figure 4.16 depicts how audio signals are distributed in the PCA space when projected onto the first two PCs, with each class represented by a distinct colour. Convex hulls are drawn to illustrate class boundaries and highlight the degree of separation in the PCA space.

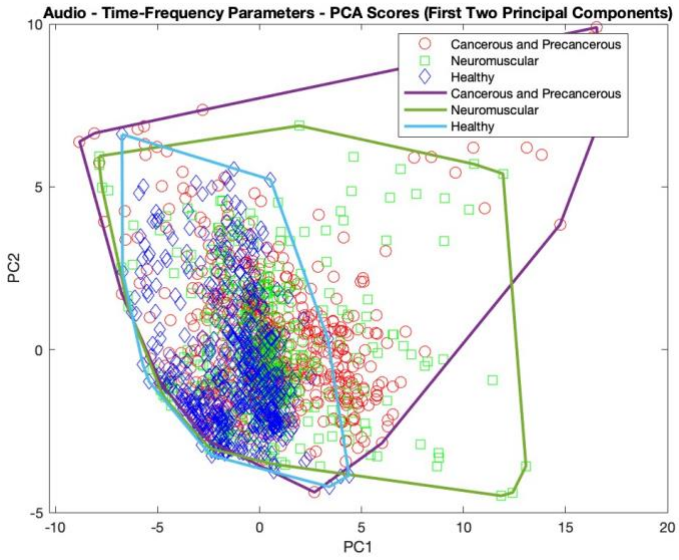

Figure 4.16: PCA score plot of Audio-derived Time-Frequency Parameters: Audio dataset of three investigated classes projected within PCA space using first two PCs derived from Time-Frequency Parameters. The boarders represent the convex hulls derived for each class in the two-dimensional space.

The Hopkins statistic computed for the first two PCs derived from the Time-Frequency Parameters calculated for the three classes of investigated audio signals indicates that the data exhibits satisfactory clustering tendencies. The corresponding H values were:

- For class of cancerous and precancerous: 0.9011,
- For class of neuromuscular: 0.7996,

- For healthy class: 0.8339.

Additionally, the pairwise Euclidean distances between the class centroids in the PCA space were:

- Cancerous and precancerous vs Neuromuscular: 0.5490,
- Cancerous and precancerous vs Healthy: 2.3020,
- Neuromuscular vs Healthy: 1.7841.

4.3.3.2. Bioimpedance-based PCA

The PCA performed on Time-Frequency features derived from laryngeal bioimpedance signals further improved the class separability between healthy and pathological cases, with spectral spread and spectral roll-off point enhancing their differentiation. The slight separability between the two pathological classes is also visible, however, the strong overlapping tendencies remained prevalent. Additionally, the pairwise Euclidean Distances between class centroids in the PCA space have increased, suggesting slightly better class separability of bioimpedance-derived Time-Frequency Parameters to those derived using the Global Statistical Features approach. This result aligns with the conclusion that laryngeal bioimpedance may efficiently capture distinctive features relevant for the detection of a pathology, but they are not sufficient for the precise differentiation between the pathology types.

Figure 4.17 presents the bioimpedance signals of the three investigated classes mapped onto the first two PCs, with each class represented by a different colour. Convex hulls enclose the data points for each group, indicating their boundaries and illustrating the degree of separation in this reduced-dimensional space.

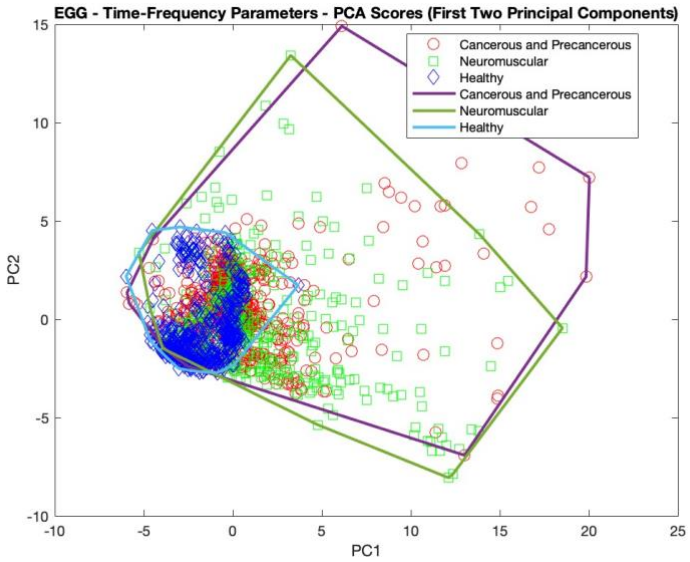

Figure 4.17: PCA score plot of EGG-derived Time-Frequency Parameters: Bioimpedance dataset of three investigated classes projected within PCA space using first two PCs derived from Time-Frequency Parameters. The boarders represent the convex hulls derived for each class in the two-dimensional space.

The Hopkins statistic computed for the first two PCs derived from the Time-Frequency Parameters of laryngeal bioimpedance signals across the three investigated classes indicates clear clustering behaviour among the laryngeal bioimpedance signals. The respective values of H were:

- For class of cancerous and precancerous: 0.9247,
- For class of neuromuscular: 0.8668,
- For healthy class: 0.8865.

Furthermore, the pairwise Euclidean distances between the class centroids in the PCA space were:

- Cancerous and precancerous vs Neuromuscular: 1.2502,
- Cancerous and precancerous vs Healthy: 2.3575,
- Neuromuscular vs Healthy: 3.0628.

4.3.4 Investigation of Class Separability using Combined Parameters and PCA

Based on the previous sections, an overall conclusion can be drawn that neither Global Statistical Feature set or the Time-Frequency Parameters deliver sufficient information for an accurate differentiation between all three classes of signals investigated in this research – cancerous and precancerous, neuromuscular, and healthy. Nevertheless, both parameter sets deliver valid information, helpful in distinguishing between the pathological and healthy cases. To ensure a comprehensive overview of the custom dataset's class separability, we performed a combined PCA on the concatenated z-scored parameter sets: the Global Statistical Features and Time-Frequency Parameters for audio, and the Global Statistical Features and Time-Frequency Parameters for bioimpedance.

Alike in the above methods, the first two PCs were used to visualise clustering tendencies among the investigated classes of both data modalities (Figure 4.20 and Figure 4.21). Below, a Pareto bar chart was used to depict the total explained variance of the first 10 PCs for both audio and bioimpedance (Figure 4.18).

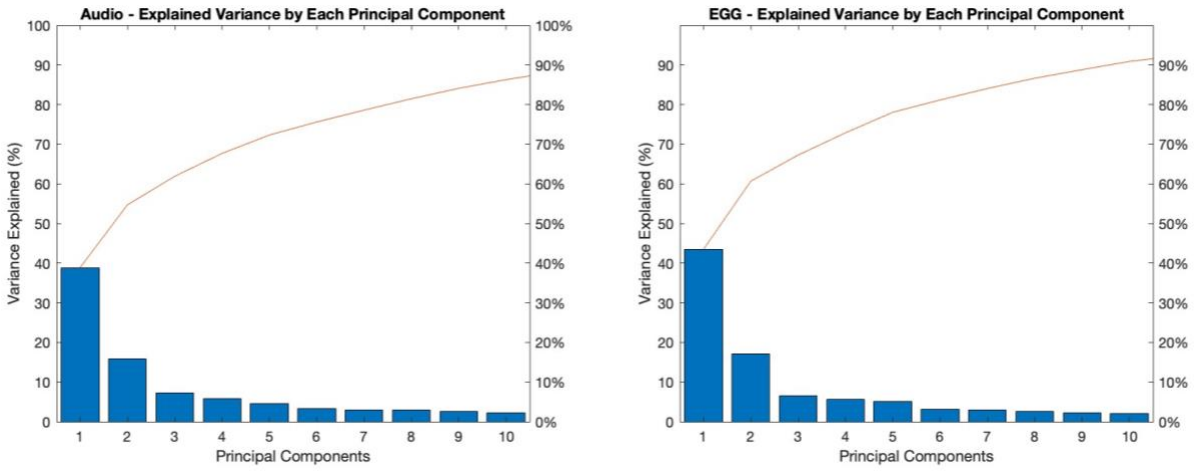

Figure 4.18: Pareto bar chart of explained variance of the first 10 principal components obtained for the combined feature set derived from both data modalities – audio and laryngeal bioimpedance (EGG).

The contribution of the derived features to the first PC obtained in the combined analysis approach varied depending on the data modality. For both audio and bioimpedance,

features contributing the least to the first PC included SD and mean (due to the z-score normalisation of the data), as well as skewness and THD. Conversely, spectral flatness (both its mean and the SD), spectral spread, and spectral entropy, as well as the mean value obtained from the STFT magnitude spectrograms were among the features with the largest contribution to the first PC. Figure 4.19 illustrates the contributions of the combined feature set to the first PC.

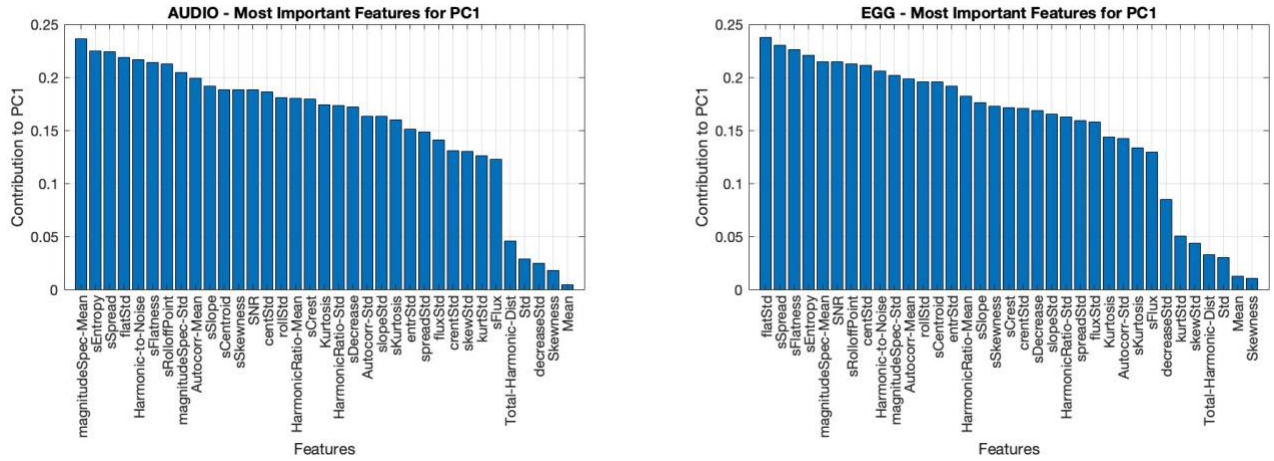

Figure 4.19: Contribution of the parameters from the Combined Feature set to the first principal component for both modalities. "Std" stands for standard deviation (SD), "Autocorr-Mean" and "Autocorr-Std" stand for mean and SD values of autocorrelation features, "SNR" stands for signal to noise ratio, "magnitudeSpec-Mean" and "magnitudeSpec-Std" signifies respectively mean and SD of magnitude spectrograms, "Total-Harmonic-Dist" stands for the total harmonic distortion. The lower "s" preceding the name of the parameters stands for "spectral", while "Std" succeeding the parameters' name stands for standard deviation, where "flat" is spectral flatness, "cent" is spectral centroid, "roll" is spectral roll-off point, "spread" is spectral spread, "entr" is spectral entropy, "slope" is spectral slope, "flux" is spectral flux, "crest" is spectral crest, "skew" is spectral skewness, "kurt" is spectral kurtosis, and "decrease" is spectral decrease.

4.3.4.1. Audio-based PCA

The PCA performed on audio-derived combined features demonstrated further improvement in class separability compared to either set of features used on audio in isolation (Global Statistical or Time-Frequency), especially for differentiation between healthy and pathological cases – those remained distinctly separated, while the overlap between the two types of pathologies (malignant and neuromuscular) persisted. In comparison with PCA analysis of audio-derived parameters obtained for either of the parameter sets (either Global Statistical or Time-Frequency), the pairwise Euclidean

Distances between the centroids of all three classes increased, indicating the combination of two types of features improved the overall class separability of the custom dataset. Nevertheless, those values did not exceed the pairwise Euclidean Distances calculated for the bioimpedance signals in Time-Frequency Parameter approach.

The following figure (Figure 4.20) presents the distribution of audio data projected onto the PCA space using the first two coefficients. The convex hulls enclose the data points for each class, illustrating their boundaries and the degree of class separation.

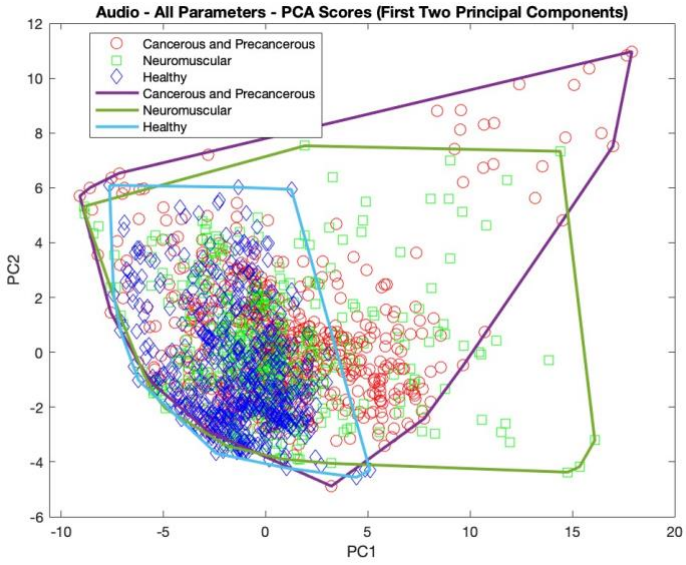

Figure 4.20: PCA score plot of Audio-derived Combined Parameters: Audio dataset of three investigated classes projected within PCA space using first two PCs derived from the Combined Parameters. The boarders represent the convex hulls derived for each class in the two-dimensional space.

The Hopkins statistics H computed for the first two PCs of each class were:

- For class of cancerous and precancerous: 0.8644,
- For class of neuromuscular: 0.8421,
- For healthy class: 0.8363.

The pairwise Euclidean distances between the class centroids in the PCA space were:

- Cancerous and precancerous vs Neuromuscular: 0.7813,
- Cancerous and precancerous vs Healthy: 2.7721,

- Neuromuscular vs Healthy: 1.9948.

4.3.4.2. Bioimpedance-based PCA

The PCA applied to laryngeal bioimpedance signals using the combined feature approach revealed greater separability for pathological and healthy cases, particularly along the PC1, reinforcing the potential of EGG for pathology detection. With neuromuscular and malignant conditions continuing to show a degree of overlap, the class separation clearly improved compared to using either parameter set in isolation. This is evidenced with the increased pairwise Euclidean Distances between classes' centroids, which were considerably larger for the combined analysis of bioimpedance signals than for any other data or feature case.

Figure 4.21 displays the bioimpedance signals projected onto the first two PCs, with each class distinguished by a colour-coded convex hulls, visually highlighting the class boundaries in the PCA space.

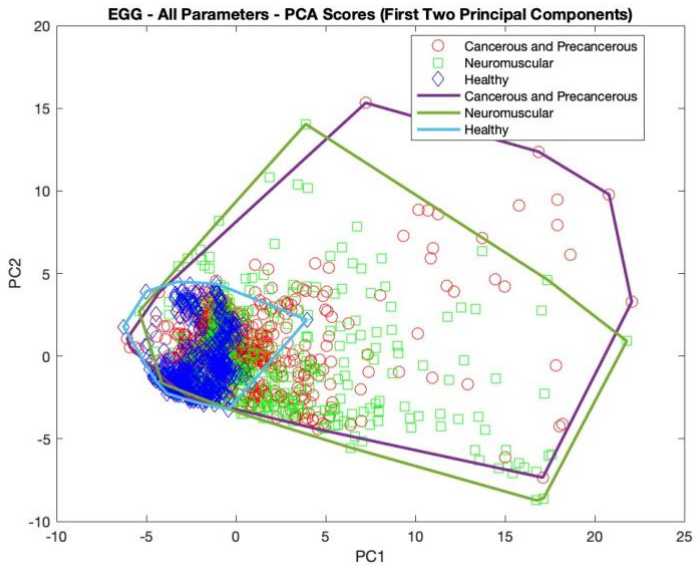

Figure 4.21: PCA score plot of EGG-derived Combined Parameters: Bioimpedance dataset of three investigated classes projected within PCA space using first two PCs derived from the Combined Parameters. The boarders represent the convex hulls derived for each class in the two-dimensional space.

The Hopkins statistic H computed for the first two PCs of laryngeal bioimpedance signals confirmed strong clustering behaviour within the data:

- For class of cancerous and precancerous: 0.9450,
- For class of neuromuscular: 0.8668,
- For healthy signals: 0.8976.

The Pairwise Euclidean Distances between class centroids for EGG signals in the PCA space were as follows:

- Cancerous and precancerous vs Neuromuscular: 1.4938,
- Cancerous and precancerous vs Healthy: 2.8362,
- Neuromuscular vs Healthy: 3.8053.

4.3.5 Combined Multimodal Analysis of Class Separability using PCA

The integration of multiple data modalities can enhance the performance of a classification system, also in case of laryngeal events (Miliaresi *et al.*, 2022; Muhammad and Alhussain, 2021; Geng *et al.*, 2022). Given that laryngeal pathologies can manifest through both bioimpedance as well as acoustic signals, combining the two data modalities has the potential of improving the pathology differentiation. Following this hypothesis, the last stage of the data analysis entailed investigation of the class separability through the analysis of parameters derived from both audio signals and the laryngeal bioimpedance and concatenated into a unified feature space.

This analysis aimed to establish whether the integration of both modalities indeed improves pathology differentiation beyond what is achievable using each modality separately. For that, PCA was conducted on the concatenated combined parameters set, incorporating both Global Statistical Features and Time-Frequency Parameters extracted from both audio and bioimpedance signals.

The first two principal components (PC1 and PC2) were examined to visualise clustering tendencies among the investigated classes (Figure 4.24). Figure 4.22 presents a Pareto bar chart illustrating the variance percentage explained by the first 10 principal components derived from the concatenated parameters of both data modalities – the multimodal approach.

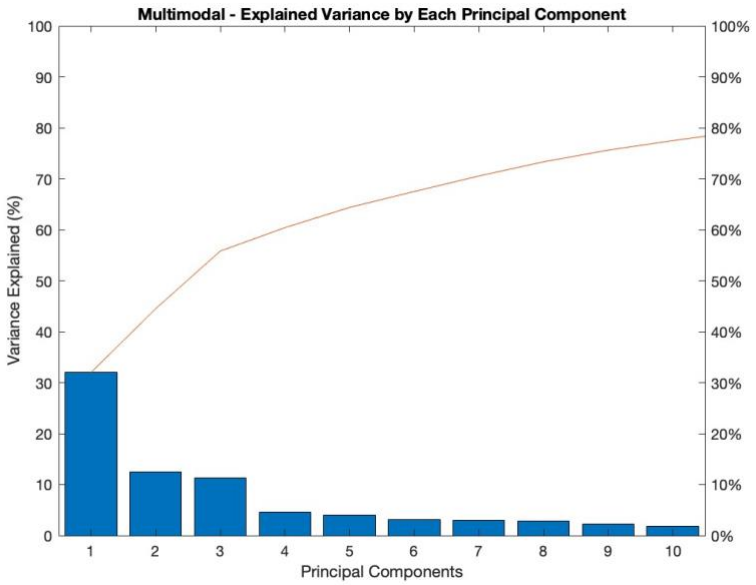

Figure 4.22: Pareto bar chart of explained variance of the first 10 principal components obtained for the multimodal feature set – concatenated parameters derived from both audio and laryngeal bioimpedance (EGG).

When analysing the combined feature set of concatenated parameters derived from audio and bioimpedance – the multimodal feature set – the most impactful features contributing to PC1 included parameters derived from the laryngeal bioimpedance; spectral flatness (SD and mean), harmonic ratio (mean), spectral centroid (SD), and spectral entropy (SD). Those were followed by spectral flatness parameter (SD and mean) calculated for the audio signals. These results indicate that the bioimpedance-derived spectral features play a dominant role in distinguishing between the classes. Figure 4.23 presents the contributions of the derived parameters to the first PC in a form of a bar chart.

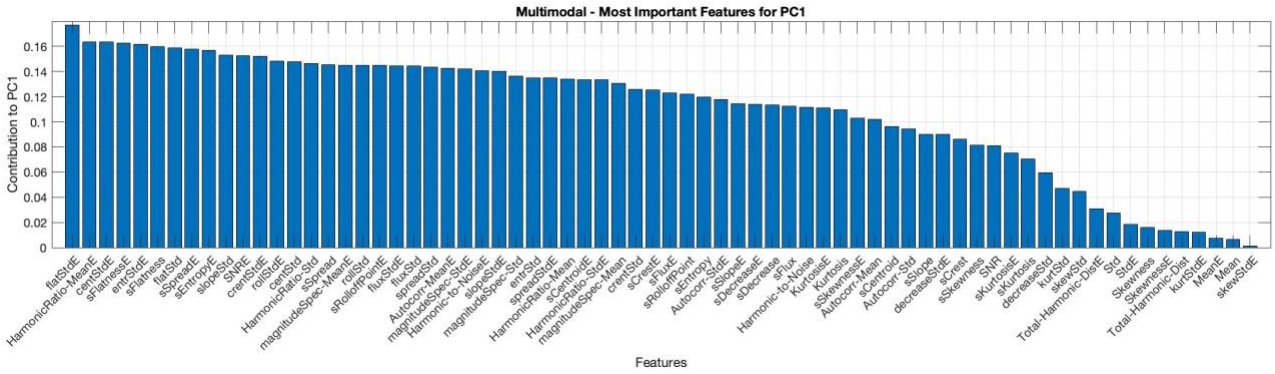

Figure 4.23: Contribution of the Multimodal Features (concatenated combined parameters derived from audio and bioimpedance) to the first principal component. All parameters ending with "E" signify features derived from the bioimpedance signals, where "Std" stands for standard deviation (SD), "Autocorr-Mean" and "Autocorr-Std" stand for mean and SD values of autocorrelation features, "SNR" stands for signal to noise ratio, "magnitudeSpec-Mean" and "magnitudeSpec-Std" signifies respectively mean and SD of magnitude spectrograms, "Total-Harmonic-Dist" stands for the total harmonic distortion. The lower "s" preceding the name of the parameters stands for "spectral", while "Std" succeeding the parameters' name stands for standard deviation, where "flat" is spectral flatness, "cent" is spectral centroid, "roll" is spectral roll-off point, "spread" is spectral spread, "entr" is spectral entropy, "slope" is spectral slope, "flux" is spectral flux, "crest" is spectral crest, "skew" is spectral skewness, "kurt" is spectral kurtosis, and "decrease" is spectral decrease.

The PCA performed on the multimodal feature set (Global Statistical parameters and Time-Frequency parameters derived from audio and bioimpedance signals, concatenated in a unified feature set) revealed enhanced class separability compared to either parameter set (or combined parameters) obtained in unimodal approaches. However, a degree of overlap remained between neuromuscular and cancerous conditions, suggesting that while a multimodal approach improves class differentiation, some pathologies still exhibit similar signal characteristics. While the Hopkins statistic remained comparable to those obtained in previous analyses, the Euclidean Distances increased considerably, indicating that the multimodal approach of combining audio and laryngeal bioimpedance improved pathology differentiation.

The following figure (Figure 4.24) demonstrates the distribution of multimodal features projected onto the PCA space using the first two principal components, with convex hulls highlighting classes' boundaries, representing present clustering tendencies and the degree of class separation.

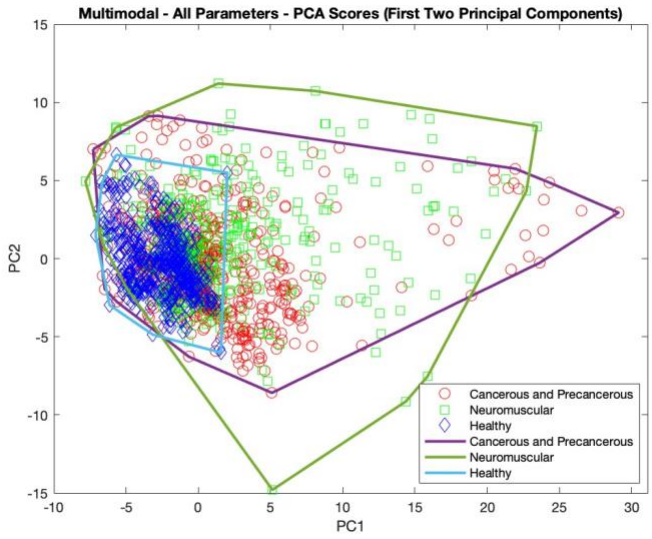

Figure 4.24: PCA score plot of Multimodal Features: multimodal dataset of three investigated classes projected within PCA space using first two PCs derived from the combined parameters concatenated in the multimodal approach. The boarders represent the convex hulls derived for each class in the two-dimensional space

The Hopkins statistic computed for the first two PCs of each class were:

- For class of cancerous and precancerous: 0.8775,
- For class of neuromuscular: 0.8647,
- For healthy class: 0.8215.

Additionally, the pairwise Euclidean distances between the three class centroids were computed in the PCA space:

- Cancerous and precancerous vs Neuromuscular: 1.5764,
- Cancerous and precancerous vs Healthy: 3.9591,
- Neuromuscular vs Healthy: 4.2861.

4.3.6 Summary of Findings and Implications for Model Development

The preliminary classification experiments revealed several critical insights:

- Speaker-dependent features can significantly inflate classification accuracy – a thorough investigation of laryngeal pathology classification requires a strict participant-independent dataset division.
- Some laryngeal pathologies cannot be classified as separate conditions due to their symptom overlap.
- A three-class classification approach (including the cancerous and precancerous growths, neuromuscular disorders, and healthy cases) aligns with clinical needs, resulting in more reliable and interpretable classification outcomes.

Due to these reasons, the dataset was refined and restructured into three major classes of laryngeal condition, with participant-independent partitioning in training and validation sets for classification purposes, ensuring that the final system was designed with clinical applicability and diagnostic relevance in mind.

Following the restructuring of the final dataset, the exploratory data analysis was pursued, focusing on the exploration of statistical and spectral parameters derived from audio, as well as bioimpedance signals. The multimodal approach was also investigated by concatenating the parameters obtained from both data modalities into a unified feature set. All parameter sets were examined using PCA, from which Hopkins statistics (presented in Table 4.7) and Euclidean Distances (presented in Table 4.8) were calculated to examine the class separability. The following tables depict the Hopkins statistics calculated for each data setup (audio and bioimpedance separately, as well as a multimodal approach) in all parameters sets, as well as the pairwise Euclidean Distances calculated for the centroids of each investigated class of signals:

Table 4.7: Hopkins statistics calculated for PCA performed on three investigated classes of laryngeal pathologies – parameters calculated for each data modality and each feature set approach.

	Cancerous and precancerous	**Neuromuscular**	**Healthy**
Global Statistical Features – Audio	0.9500	0.8910	0.8240
Global Statistical Features - Bioimpedance	0.9592	0.8319	0.8331
Time-Frequency Parameters – Audio	0.9011	0.7996	0.8339
Time-Frequency Parameters – Bioimpedance	0.9247	0.8668	0.8865
Combined Parameters – Audio	0.8644	0.8421	0.8363
Combined Parameters – Bioimpedance	0.945	0.8668	0.8976
Combined Parameters – MULTIMODAL APPROACH	0.8775	0.8647	0.8215

Table 4.8: Euclidean Distances between class centroids of three investigated classes of laryngeal pathologies – parameters calculated for each data modality and each feature set approach.

	Cancerous and precancerous vs Neuromuscular	**Cancerous and precancerous vs Healthy**	**Neuromuscular vs Healthy**
Global Statistical Features – Audio	0.6201	1.5583	0.9426
Global Statistical Features - Bioimpedance	0.7361	1.5811	2.2548
Time-Frequency Parameters – Audio	0.5490	2.3020	1.7841
Time-Frequency Parameters – Bioimpedance	1.2502	2.3575	3.0628
Combined Parameters – Audio	0.7813	2.7721	1.9948
Combined Parameters – Bioimpedance	1.4938	2.8362	3.8053
Combined Parameters – MULTIMODAL APPROACH	1.5764	3.9591	4.2861

The exploratory data analysis revealed the following:

- PCA performed on Global Statistical Feature set and Time-Frequency Parameters show significant differences between healthy and pathological signals for either

data modality, with a slight overlap present. This indicates the development of a laryngeal pathology detection system is possible, however, additional application of more sophisticated feature extraction methods and deep learning classification can be harnessed to improve the system's performance and accuracy.

- Either set of parameters alone are insufficient for accurate multi-class laryngeal pathology classification – distinct clustering tendencies are visible, however, a predominant overlap between the two pathologies is present. This reinforces the need for the advanced feature extraction and application of deep learning-based classification.
- Spectral features play a crucial role in separating laryngeal pathology groups – the combined analysis showed features including spectral flatness, spectral entropy, spectral spread and the parameters derived from magnitude spectrograms contribute significantly more to PC1 than any of the time-domain features.
- Bioimpedance-based PCA provided a clearer separation between all classes than audio-based PCA – especially for healthy and pathological cases – which was evidenced using pairwise Euclidean Distances. This reinforced the potential of EGG in laryngeal pathology classification.
- The combined feature approach resulted in greater between-class Euclidean Distances for both data modalities, indicating the integration of multiple feature representations enhances pathology differentiation.
- The multimodal approach combining features of both data modalities significantly enhances the class separability, as PCA plots exhibit larger Euclidean Distances and clustering patterns than those obtained from a single modality.

These findings highlighted the potential of digital signal-based laryngeal pathology classification, with an emphasis on the capabilities of laryngeal bioimpedance. Furthermore, the findings support the hypothesis that a multimodal approach to laryngeal pathology classification outperforms unimodal approaches. However, the findings further validate the necessity of more advanced feature extraction and classification approaches for multimodal pathology classification – such as, for instance, deep-learning classification. The improved separability observed with the combined approach suggests that integrating time and frequency domains for the feature extraction will be beneficial in subsequent deep learning-based classification experiments.

4.4. SAARBRUECKEN VOICE DATABASE OVERVIEW

The Saarbruecken Voice Database (SVD) is the most commonly used public database of audio and simultaneously recorded EGG signals in the research of laryngeal pathology classification. Initially developed in 1997 by Manfred Pützer and Jacques Koreman in collaboration with the Department of Phoniatrics and ENT at the Caritas Clinic St. Theresia in Saarbrücken (Pützer and Koreman, 1997), it is now managed by Manfred Pützer and William J. Barry (Saarbruecken Voice Database: Handbook, 2023). Reportedly, the database contains recordings of 71 distinct vocal tract pathologies (Barry and Putzer, 2007). Each recorded participant was asked to perform three tasks during data collection: sustaining vowel phonation (vowels [i], [a], [u]) at normal, high, and low pitch; performing the same sustained vowel phonations at rising-falling pitch; and a recording of the phrase “Guten Morgen, wie geht es Ihnen?” (“Good morning, how are you?”). All recordings are sampled at 50000 Hz with 16-bit depth resolution.

4.4.1 Limitations of SVD

It has been reported that SVD contains recordings from over 2000 speakers (Zhou *et al.*, 2022; Lee, 2021; Barry and Putzer, 2007), of which 1356 are pathological (Harar *et al.*, 2017). However, this number does not correspond to unique individuals. A thorough investigation of SVD has revealed that some participants have been recorded multiple times – in some cases, as many as 24 times (a participant no. 2027, suffering from spasmodic dysphonia) – and have been assigned different identification numbers depending on a recording session number. In total, 336 duplicates of participants suffering from laryngeal pathologies can be found throughout the database with various ID numbers. As a result, the actual number of unique subjects within the pathological subset of SVD must be reported as 1020.

It should be noted that the presence of multiple recordings gathered from the same subjects can leverage speaker-dependent features and lead to misleadingly high classification accuracy results, as previously noted (*4.2. Preliminary Investigation of Custom Dataset Classification*). To maintain participant independence throughout the training and validation processes, it must be ensured that recordings belonging to the same subject do not appear in both the training and validation datasets. Otherwise, such an error could introduce speaker-dependent bias into the classification process, resulting in artificially elevated system validation performance.

Another limitation posed by SVD is the presence of conditions not typically considered as pathological from the medical perspective – among 1020 participants from the pathological group of SVD, some have been assigned the conditions referred within SVD as:

- “Gesangsstimme” (17 participants) and “Sengerstimme” (2 participants): both terms relate to the voice of a professional singer, therefore neither is a pathology.

- “Mutatio” (2 participants): refers to normal voice changes during puberty.
- “Morbus Down” (2 participants): Down Syndrome, a genetic disorder.
- “Morbus Parkinson” (1 participant): Parkinson’s Disease, a neurological disorder.
- “Vox senilis” (40 participants): Changes in voice due to aging, not a pathological condition.

A total of 41 participants were assigned one of the conditions listed above as their sole diagnosis. After excluding these cases, the pathological subset of SVD comprises 979 unique participants with laryngeal pathologies.

Medical notes provided in SVD indicate that at least 53 of the remaining participants from pathological subset had undergone treatment prior to data collection (e.g., participants with IDs 1454, 1475, 1743). These cases should be interpreted with caution, as their recordings may no longer contain features representative of the original pathology category.

Additionally, 21 participants from remaining 979 were under 18 years old. Since this study focuses exclusively on adults, these participants were excluded, reducing the final pathological subset to 958 unique adult participants.

For the healthy group, while 869 participants are listed (Barry and Putzer, 2007), 36 are under 18, and 15 are duplicates (i.e., the same individuals assigned different ID numbers). After accounting for these exclusions, the healthy subset consists of 818 unique adult participants.

These elements have been considered significant limitations of SVD. In this study, we address a critical issue in laryngeal pathology classification specifically concerning the independence of training and validation datasets to ensure robust and generalisable models. Previous research that relies on SVD often reports high classification accuracies; however, it often lacks clarification on whether the models are speaker-independent, which is essential for reliable clinical applications. SVD contains recordings of the same

participants under different IDs, which – unless addressed appropriately – can lead to inflated accuracy due to the classifier's ability to recognise the subject's voice. To mitigate this, for the purposes of this research, we assigned a new identification number to each participant, and meticulously selected samples to ensure complete speaker independence, assigning each participant's recordings exclusively to either the training or validation dataset. By doing so, we prevent the classifier from leveraging participant-specific characteristics across datasets. Overall, it would be recommended that work using SVD provides clarification as to the participant-independence of the reported models.

4.4.2 SVD Participant Selection

To develop a robust laryngeal pathology classification system, two distinct classification tasks were defined. The first aimed to detect the presence of a laryngeal pathology regardless of its aetiology (binary pathology detection system), while the second focused on distinguishing between three specific pathology categories: precancerous and cancerous growths, neuromuscular disorders, and healthy individuals. To meet the requirements of both systems, two different subsets of SVD were selected. In both cases, to eliminate the bias caused by the inclusion of the same participants in both training and validation datasets, a smaller subset of SVD was selected.

4.4.2.1. Multi-class Classification

For the purposes of the multi-class classification system, where the primary objective was developing a laryngeal pathology classification system capable of distinguishing between three categories – precancerous and cancerous growths, neuromuscular disorders, and healthy individuals, a more refined subset of SVD was selected.

For the precancerous and cancerous class, the following conditions were selected from SVD; Hypopharyngeal Tumour ("Hypopharynxtumor" – A), Laryngeal Tumour

(“Kehlkopftumor” – B), Laryngeal Pachydermia (“Kontaktpachydermie” – C), Leukoplakia (“Leukoplakia” – D), Vocal Cord Carcinoma (“Stimmlippenkarzinom” – E). Certain cancerous and precancerous pathologies, such as Carcinoma in Situ, Epiglottic Carcinoma ("Epiglottiskarzinom"), Mesopharyngeal Tumour ("Mesopharynxtumor"), and Papilloma ("Papillom"), were excluded due to an insufficient number of recordings, with some classes containing data from only one participant. Furthermore, recordings for some of these cases were made post-treatment, rendering them unsuitable for this study.

From the initial 77 participants in the precancerous and cancerous subset, 16 were duplicates, and 7 had undergone surgical treatment prior to data collection. After these exclusions, the usable dataset for this class comprised 54 unique participants. Individual counts for specific conditions are presented in Figure 4.25, where "A" represents Hypopharyngeal Tumour, "B" Laryngeal Tumour, "C" Laryngeal Pachydermia, "D" Leukoplakia, and "E" Vocal Cord Carcinoma.

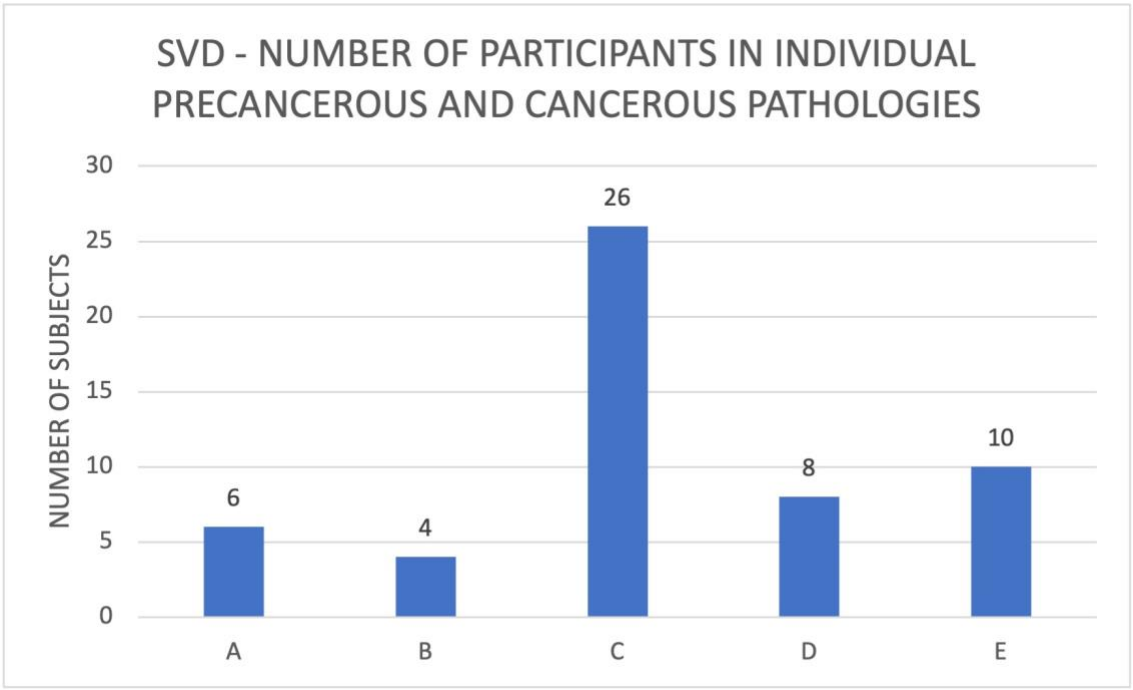

Figure 4.25: Participant numbers for each cancerous and precancerous condition selected from SVD.

For the neuromuscular disorder class, Recurrent Laryngeal Nerve Palsy ("Rekurrensparese" in SVD) was chosen. After excluding duplicate participants and individuals under the age of 18, this subset contained 144 unique participants.

Among the 818 healthy participants in SVD, 200 were randomly selected for each cross-validation run to reduce dataset imbalance. The final number of participants in each category is presented in Figure 4.26.

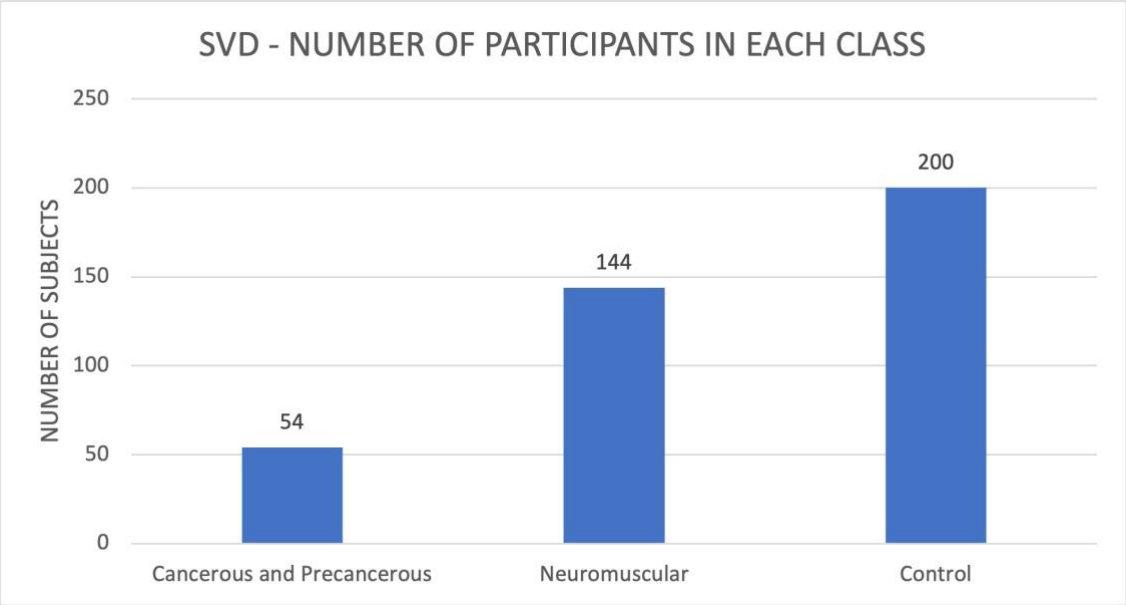

Figure 4.26: Number of participants in each category in SVD.

4.4.2.2. Binary System – Detection of Laryngeal Pathology

The purpose of the binary classification system was to detect the presence of any laryngeal pathology, regardless of its aetiology. To ensure the developed system's capability of detecting the presence of a pathology even if unrelated to vocal fold paralysis or malignant lesions, a broader pathological subset of SVD was chosen. This subset included the pathological conditions used for the multi-class system (hypopharyngeal tumour, laryngeal tumour, laryngeal pachydermia, leukoplakia, vocal cord carcinoma, as well as the recurrent laryngeal nerve palsy) as well as the additional pathologies including laryngitis, phonation nodules (in SVD – "Phonationsknotchen"), Reinke's oedema, and vocal cord polyps

(“Stimmlippenpolyp”). In total, 10 different laryngeal disorders from SVD were used for the binary classification system. This included the total of 362 participants affected by laryngeal pathologies. To match this number, the subset of healthy subjects was broadened to 400 to account for any missing samples after the pre-segmentation stage of data preprocessing.

This approach ensured that the developed system could generalise across a variety of laryngeal pathologies, detecting the presence of disease even when the specific pathology was unrelated to vocal fold paralysis or malignant lesions.

4.4.3 SVD’s Data Preprocessing

The preprocessing of SVD was conducted using the same methodology as the latter stage of preprocessing applied to the custom dataset. This approach ensured consistency between the two datasets, enabling accurate comparisons and seamless integration for classification tasks.

The preprocessing was implemented using the MATLAB computing environment and consisted of the following stages:

1. Sample rate change:

As described at the start of this section (*4.4. Saarbruecken Voice Database*), all signals obtained from SVD were captured at a sampling rate of 50000 Hz. To ensure consistency between SVD and the custom datasets, allowing for the application of the same feature extraction and classification algorithms without introducing discrepancies caused by differing sampling rates, the sampling rate change was the first and most crucial step in SVD signal preprocessing. All audio signals were resampled from the initial 50000 Hz to 44100 Hz to match the sampling rate used in the custom dataset.

All laryngeal bioimpedance signals obtained from SVD were also resampled accordingly. Following the approach taken in the latter stage of the signal preprocessing conducted for

the custom dataset, the SVD signals were first subjected to the antialiasing low-pass filtering, followed by resampling from the initial 50000 Hz down to 2048 Hz. This reduced sampling rate was sufficient to retain diagnostically significant features while greatly reducing computational overhead.

2. Band-pass filtering of audio:

Similarly to preprocessing of the custom dataset, the audio signals were subsequently subjected to a band-pass filtering with the frequency range of 50-17000 Hz. This was achieved using two IIR filters designed for the implementation of band-pass filtering of the custom dataset (for details, see *4.1.3 Data Preprocessing* section).

3. Amplitude normalisation:

To address the issue of signal standardisation, z-score normalisation was applied to both audio and bioimpedance signals immediately prior to the feature extraction stage. This normalisation technique ensured consistent scaling across the dataset. Peak normalisation, which had previously been tested, was excluded from preprocessing as it was found to inadequately address sudden amplitude changes and offered less robust standardisation compared to z-score.

4. Data pre-segmentation:

Following the final stage of the custom dataset preprocessing, the recordings obtained from SVD were also subjected to the segmentation stage. To standardise the data for analysis, all recordings were split into one-second intervals, with each segment corresponding to 44100 samples for audio data and 2048 samples for the bioimpedance. The intervals were then saved as separate mono WAV files in folders corresponding to their pathology categories.

After performing all data preprocessing stages on the SVD data, including the pre-segmentation stage, the total number of data samples was as follows (Table 4.9 and 4.10):

Table 4.9: Number of data samples obtained from SVD in each category (after data preprocessing stage) and used in the multi-class laryngeal pathology discrimination system.

Signal Subset of SVD	**Audio Sustained Phonation Dataset**	**Audio Continuous Speech Dataset**	**Bioimpedance Sustained Phonation Dataset**	**Bioimpedance Continuous Speech Dataset**
Cancerous and Precancerous	617	142	617	142
Neuromuscular	1124	351	1124	351
Control (Healthy)	2468	568	2468	568
TOTAL PATHOLOGICAL	1741	493	1741	493

Table 4.10: Number of data samples obtained from SVD in each category (after data preprocessing stage) and used in the pathology detection system (binary between healthy and pathological).

Signal Subset of SVD	**Audio Sustained Phonation Dataset**	**Audio Continuous Speech Dataset**	**Bioimpedance Sustained Phonation Dataset**	**Bioimpedance Continuous Speech Dataset**
Pathological	2865	766	2865	766
Control (Healthy)	4910	771	4910	771

This preprocessing pipeline ensured that the SVD signals were prepared in a manner consistent with the custom dataset, allowing for a robust and fair comparison during classification evaluation.

Chapter 5

Feature Extraction Methods

The performance of a classification system relies majorly on the quality and the form of the data delivered to the classifier. It is especially important for machine-learning and deep-learning-based algorithms, particularly for tasks related to diagnostics of pathological states. Since the crucial objective of this research is to optimise the classification of audio and laryngeal bioimpedance signals based on the health condition of the larynx, we apply several feature extraction methods to assess their performance and choose most suitable ones for each data modality.

In this chapter, we first review the basic forms of audio and bioimpedance signals in both time and frequency domains as the fundamental signal analysis methods (section 5.1) – the WAV files showcasing signal's waveform, and the STFT spectrogram representing its frequency spectrum. In the following sections (5.2 and 5.3) we explore perceptually and biologically inspired frequency representations, respectively, Mel and Equivalent Rectangular Bandwidth (ERB). Following the introduction of the Mel filter bank and the Gammatone filter bank, their respective spectrogram representations are discussed (Mel-spectrograms and Gammatone spectrograms), as well as their corresponding cepstral coefficients (MFCCs and GTCCs).

Summarising, this chapter discusses the feature extraction methods investigated and derived in this study. All described features are used in the subsequent stages of the research in the developed laryngeal pathology detection and classification system in order to assess their performance and choose the most appropriate methods for the final laryngeal pathology detection and classification system.

5. FEATURE EXTRACTION METHODS

5.1. TIME VS FREQUENCY REPRESENTATION

The analysis of audio signals can be approached from two fundamental perspectives: time domain and frequency domain. The time domain version of a signal can simply be represented with a plot of the WAV file recording of a signal, while its frequency domain equivalent can be calculated using STFT and represented in a form of a spectrogram. Both representations constitute primary ways to analyse recordings of laryngeal conditions. As such, these representations serve as the foundation for various sophisticated feature extraction algorithms in speech and signal analysis.

Since both methods of signal representation – the WAV files waveform and a spectrogram – form the basis for more advanced feature extraction techniques, such as cepstral coefficients and auditory-modelling representations, the primary investigation of these methods is essential. Basic time- and frequency-domain representation of a signal allows for an initial assessment of signal's quality and characteristics, which is particularly important in tasks related to pathology diagnostics.

5.1.1 Signals in Time Domain – WAV files

In the time domain, a digital signal can be represented as a waveform, where the horizontal axis corresponds to the independent variable of time – for instance, measured in samples – and the vertical axis corresponds to the signal's amplitude – representing, for example, the fluctuations of air pressure expressed as voltage. A digital signal such as audio recording of human voice or bioimpedance fluctuations measured over time, can easily be recorded and stored in a form of a Waveform File Format – a WAV file. A WAV file is a standard audio file format used for storing uncompressed signals digitally.

For the purposes of this study, all data – both audio and laryngeal bioimpedance – was recorded in a form of multiple mono WAV files, originally with 44100 Hz sampling rate and the bit depth of 16 bits per sample. In the latter stage of data preprocessing, due to the limited initial frequency content of electroglottographic measurements, the sampling rate of the bioimpedance signals was converted to 2048 Hz. All data was then segmented into 1-second-long signals. The following figure (Figure 5.1) shows an example of the final WAV file representation of simultaneous audio and bioimpedance signals recorded from a healthy participant during continuous speech:

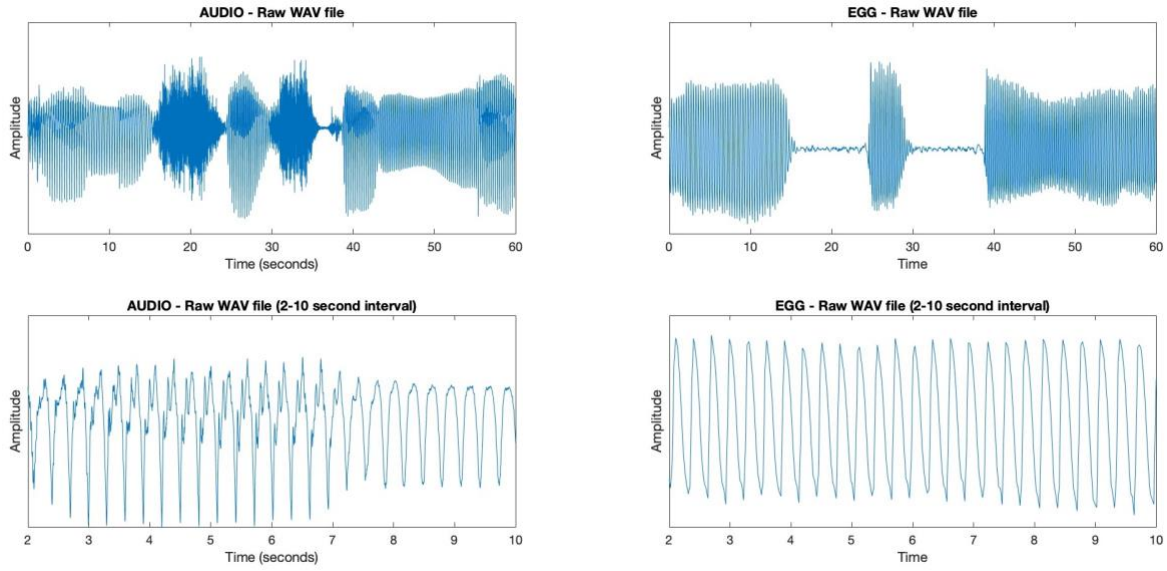

Figure 5.1: WAV file representation of an audio signal (left) and a laryngeal bioimpedance signal (right) recordings of continuous speech obtained from a healthy individual

5.1.2 Signals in Frequency Domain – Short-Time Fourier Transform

To analyse how the frequency content of a signal changes over time – thus, to perform its frequency-domain analysis – the Short Time Fourier Transform (STFT) is used, as per equation 4.3. The STFT divides the signal into short overlapping segments using a sliding windowing function, applies the Fourier transform to each segment, and stacks the resulting spectra over time.

Common window functions include Hamming, Hann, and Blackman windows, each balancing frequency resolution and sidelobe suppression. The choice of window length introduces a fundamental time-frequency resolution trade-off; shorter windows provide better time resolution but worse frequency resolution, and the longer windows improve frequency resolution but smear temporal details.

The spectrogram – the visual representation of the signal's energy across the time and frequency bins – is then obtained by taking the squared magnitude of the STFT:

$$S(t, f) = |X(t, f)|^2 \quad (5.1)$$

For the purposes of investigation into the feature extraction methodology of this research, the spectrograms generated using STFT function were also used. The parameterisation of those spectrograms according to the data modality is provided in the following table (Table 5.1).

Table 5.1: Parameters used for spectrogram calculation for the developed systems of laryngeal pathology detection and classification

Data Modality	**Sampling Rate**	**Window Function**	**Window Size**	**Overlap Size**
Audio	44100 Hz	Hanning	512	256
Laryngeal Bioimpedance	2048 Hz	Hanning	128	64

The spectrograms were derived from the data following its segmentation into 1-second-long recordings The below figure (Figure 5.2) represents the STFT spectrograms derived from the simultaneous audio and bioimpedance recordings of a continuous speech obtained from a healthy participant:

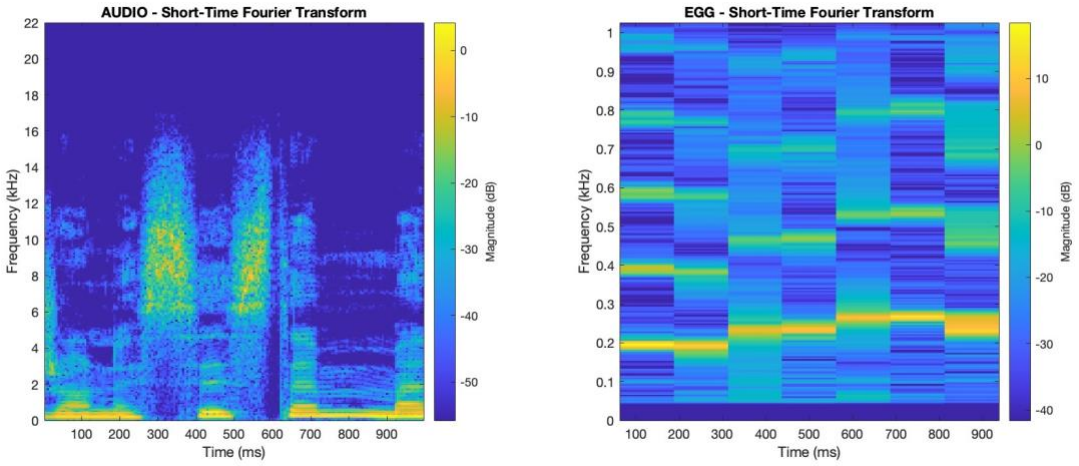

Figure 5.2: STFT spectrograms derived from an audio (left) and bioimpedance signals (right) derived from a continuous speech signal obtained from a healthy individual

5.2. MEL SPECTRUM

The Mel spectrum, first introduced by Stevens, Volkmann, and Newmann in 1937, is a perceptual scale that models how humans perceive different frequencies of sound (Stevens *et al.*, 1937; On *et al.*, 2006). The name “Mel” is derived from the word “melody”, highlighting its foundation in human pitch perception. The Mel scale constitutes the basis for modern Mel filter banks, Mel spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs), all widely used for capturing the perceptual characteristics of speech and audio signals (Imai, 1983).

The Mel spectrum divides the audible frequency range into a set of Mel-frequency bins spaced densely at lower frequencies, with sparser spacing at higher frequencies. Below 1000 Hz, the frequency bins of a Mel scale are spaced linearly, while those above 1000 Hz follow logarithmic spacing. The relationship between the physical frequency in Hz, and the perceived frequency of Mel scale can be approximated with the following equation (Molau *et al.*, 2001):

$$Mel(f_c) = 2595 \cdot \log_{10}\left(1 + \frac{f_c}{700}\right) \tag{5.2}$$

5.2.1 Mel Filter Bank

Since the Mel scale is motivated through subjective perception of frequencies, the definite mathematical equation defining the Mel filter function is highly debated and difficult to derive. Nevertheless, according to Beranek (1949), from the curves of Stevens and Volkmann the following frequency band comparison can be derived (Table 5.2):

Table 5.2: Frequency in Hertz and Mel-Frequency alignment proposed by Beranek (1949).

Hz	20	160	394	670	1000	1420	1900	2450	3120	4000	5100	6600	9000	14000
Mel	0	250	500	750	1000	1250	1500	1750	2000	2250	2500	2750	3000	3250

The Mel scale filter bank relies on the application of triangular bandpass filters, spaced along the Mel scale. The following figure (Figure 5.3) represents the Mel frequency filter bank derived using 1024-point Hamming window with an overlap of 512 points, for the frequency range of 0-4000 Hz.

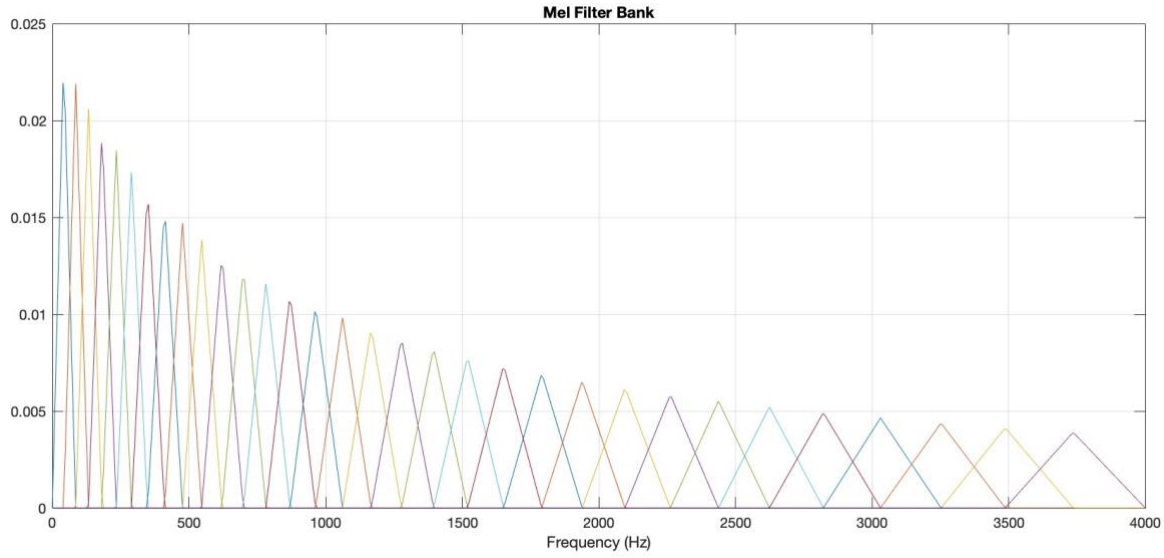

Figure 5.3: Visual representation of a Mel Scale Filter Bank

5.2.2 Mel-Spectrograms

A Mel-spectrogram is a time-frequency representation of a signal obtained by applying the Mel filter bank to the frequency spectrum of the signal derived using the STFT. It is an alternative to a regular spectrogram, that helps to visualise the frequency content of a signal over time. Mel-spectrograms can be created by calculating the magnitude spectrum of a signal using Fourier Transform and transposing the frequency spectrum into Mel scale. As an alternative to the conventional spectrogram derived using the STFT (where all frequency bins are of equal bandwidths), a Mel-spectrogram provides a frequency representation that aligns with the human-like subjective perception of audio frequencies.

The following figure (Figure 5.4) represents Mel-spectrograms derived from audio and bioimpedance recordings of continuous speech obtained from a healthy participant:

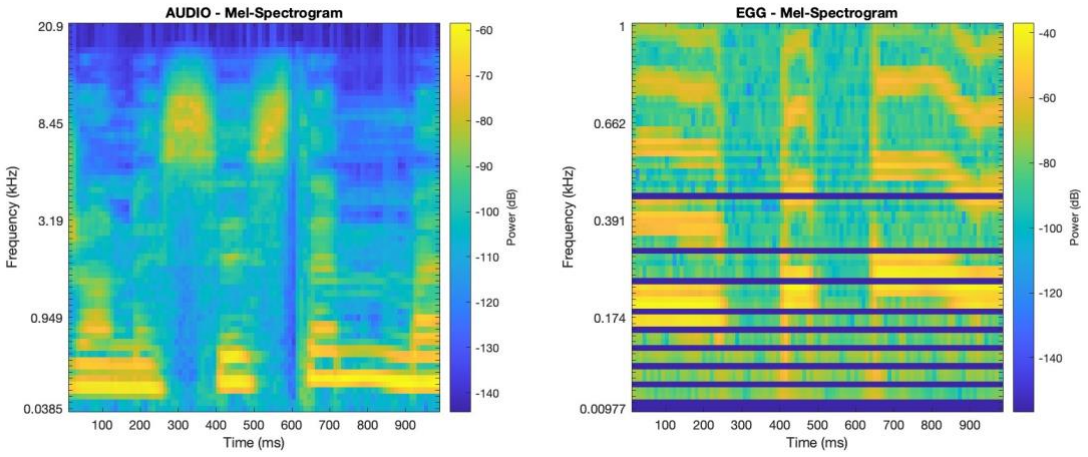

Figure 5.4: Mel-spectrograms derived from an audio (left) and bioimpedance signals (right) – a visual time-frequency representation aligned with the Mel scale derived from a continuous speech signal obtained from a healthy individual.

Mel-spectrograms are a method of visualising and analysing signals based on human subjective perception of sounds. They are particularly useful in digital signal processing for tasks such as music analysis and sound classification, with a particular emphasis on speech recognition tasks (On *et al.*, 2006; Arias-Londoño, and Godino-Llorente, 2010).

In this study, we used Mel-spectrograms as one of the feature extraction methods for the final laryngeal pathology classification model. The Mel-spectrograms were derived from the audio and bioimpedance signals preprocessed according to the methodology outlined in the previous chapter (*4.1.3. Data Preprocessing*) – 1-second-long signals, with sampling frequency of 44100 Hz for audio and 2048 Hz for bioimpedance. Each Mel-spectrogram was calculated with 64 frequency bands.

5.2.3 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCCs) are a compact representation of a signal's frequency spectrum within the Mel scale frequency range. They approximate the human perception of sound by decorrelating Mel-scaled spectral features (obtained using a Mel-spectrogram) and extracting only the most relevant information for classification tasks.

MFCCs are derived using the Discrete Cosine Transform (DCT) applied to a signal's log-scaled power spectrum aligned with the Mel scale – thus, MFCCs are obtained by applying the DCT along the frequency axis of a Mel-spectrogram (Bonet-Sola and Alsina-Pages, 2021). The detailed process of MFCC calculation entails:

1. calculation of a windowed Fast Fourier Transform (FFT, or simply the calculation of STFT),
2. application of the Mel filter bank to the magnitude spectrum of the original signal,
3. transformation of the values to their logarithmic representations by calculating the logs of the powers at each Mel frequency bin,
4. calculation of the DCT from the log-filter bank coefficients calculated in the previous step (Bonet-Sola and Alsina-Pages, 2021).

The final step of cepstral coefficient calculation involves retaining a selected number of the coefficients – usually 12 (Arias-Londoño *et al.*, 2010) – since the higher-order

coefficients capture less perceptually relevant details. Thus, the coefficients matrix resulting from the above calculations constitutes a compact representation of the essential spectral characteristics of the analysed signal.

MFCCs are particularly popular in the field of speech and audio processing due to their ability to capture essential features of sound signals in a way that aligns well with human auditory perception (Arias-Londoño *et al.*, 2010; Bonet-Sola and Alsina-Pages, 2021). In the context of speech recognition, MFCCs help in extracting relevant information from audio signals to facilitate the identification of spoken words (On *et al.*, 2006). They are characterised by their ability to represent the frequency content of a signal in a way that minimises redundancy and focuses on the most perceptually significant components of the sound. This is particularly useful in applications where high-dimensional data, such as raw audio waveforms, need to be compactly and effectively represented.

In this research, we used MFCCs as one of the feature extraction methods to be assessed by the final laryngeal pathology classification system. MFCCs were derived from both data modalities (audio and bioimpedance signals) preprocessed according to the methodology outlined in the previous chapter (*4.1.3. Data Preprocessing*). Considering the change in the sampling rate depending on the data modality (44100 Hz for audio, and 2048 Hz for laryngeal bioimpedance), the parameters for MFCCs calculation varied depending on the data type. The parameters used for the calculation of MFCCs are listed in the following table (Table 5.3):

Table 5.3: Parameters defined for MFCC calculation for the developed systems of laryngeal pathology detection and classification.

Data Modality	**Sampling Rate**	**Window Function**	**Window Size**	**Overlap Size**	**Number of Coefficients**
Audio	44100 Hz	Hanning	512	256	40
Laryngeal Bioimpedance	2048 Hz	Hanning	128	64	12

The following figure (Figure 5.5) depicts an example of the MFCC matrices derived from both data modalities for a continuous speech signal obtained from a healthy participant:

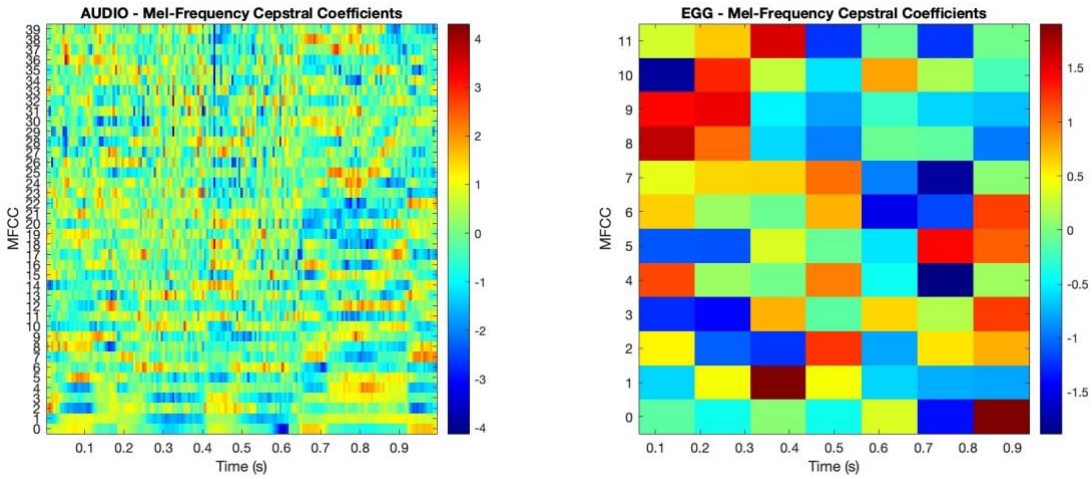

Figure 5.5: Representation of MFCCs derived from an audio (left) and laryngeal bioimpedance (right) recordings of continuous speech obtained from a healthy individual.

5.3. EQUIVALENT RECTANGULAR BANDWIDTH SPECTRUM

Equivalent Rectangular Bandwidth (ERB) is a fundamental concept in psychoacoustics, aimed at emulating the frequency sensitivity of a human ear. First referred to as “critical band” (Fletcher, 1940), the concept was introduced by Fletcher and Munson as a representation of the “auditory filter” of the basilar membrane of a human cochlea (Fletcher and Munson, 1933), which acts as a frequency analyser. The critical band was first conceptualised as the frequency bandwidth within which the perception of one tone may be interfered by the second tone through auditory masking (Fletcher, 1940). The ERB is derived based on the observation that the width of the auditory filters on the basilar membrane is approximately constant, even as the centre frequency changes. Thus, each bandwidth of the ERB spectrum corresponds to a fixed distance along the basilar membrane (Glasberg and Moore, 1990). Generally, the mathematical representation of the ERB is given by:

$$ERB(f_c) = \left[\left(\frac{f_c}{EarQ}\right)^p + (minBW)^p\right]^{\frac{1}{p}} \quad (5.3)$$

Where f_c stands for the central frequency of a filter in Hz, $EarQ$ is the asymptotic filter quality at higher frequencies, $minBW$ is the minimum bandwidth at the lower frequencies, and p is the order of approximation, typically of value 1 or 2 (Valero and Alias, 2012). Glasberg and Moore (Moore and Glasberg, 1987) suggested the ERB spectrum can be approximated by the polynomial equation:

$$ERB(f_c) = 6.23f_c^2 + 93.4f_c + 28.5 \quad (5.4)$$

and its linear representation (Glasberg and Moore, 1990; Slaney, 1993):

$$ERB(f_c) = 24.7(4.37f_c + 1) \quad (5.5)$$

5.3.1 Gammatone Filter Bank

The concept of the ERB lays the foundation for Gammatone filter banks, and thus, the Gammatone spectrograms and the Gammatone Cepstral Coefficients. The Gammatone filters are parametrised in terms of ERBs, aligning their bandwidths with the frequency bands determined by that spectrum (Patterson and Moore, 1986). In essence, the ERB spectrum informs the design of Gammatone filters, creating a connection between the perceptual aspects of human hearing and the mathematical modelling used in signal processing – while the ERB spectrum represents a psychoacoustic concept related to the perceived loudness of different frequencies, Gammatone filters are mathematical constructs designed to replicate the filtering properties of the human auditory system itself (Lyon *et al.*, 2010).

The idea of the Gammatone filter was first introduced by Johannesma (1972) as a mathematical representation of a reverse correlation – *revcor* – function (De Boer and Kuyper, 1968). The reverse correlation, investigated on cats' cochlea, was developed as an

estimate of an auditory system measuring the correlation between a white noise input stimuli with the obtained impulse response of a primary auditory nerve fibre (Darling, 1991). While the *revcor* is a continuous representation of a set of recorded data points, the Gammatone filter function is an equation invented as an analytic expression of the *revcor* phenomenon (Patterson *et al.*, 1987). The filter and its function are referred to as “gammatone”, since its impulse response $g(t)$ is a direct product of multiplication between the gammatone distribution function and a sinusoidal tone (Johannesma, 1972; Patterson *et al.*, 1987)

$$g(t) = A \cdot t^{n-1} \cdot e^{-2\pi Bt} \cdot \cos(2\pi f_c t + \varphi) \qquad \text{for} \qquad t \geq 0 \tag{5.6}$$

where:

- A is a normalisation constant responsible for the gain, usually equal to 1 (Holdsworth *et al.*, 1988),
- t represents time,
- n is the order of the filter,
- B is the frequency bandwidth of the filter,
- f_c is the centre frequency of the filter,
- φ is the initial phase shift of the filter.

This formulation explicitly shows the gamma envelope $t^{n-1}e^{-2\pi Bt}$ modulating a sinusoidal carrier $\cos(2\pi f_c t + \varphi)$. The key parameters of a Gammatone filter are n and B; B represents the duration of the filter’s impulse response, which is directly related to the ERB spectrum and represents the bandwidth of the filter. Parameter n represents the filter’s order, playing a significant role in shaping the slope of its skirts (Patterson *et al.*, 1992).

The application of a Gammatone filter function as a representation of the magnitude characteristic of a human auditory filter proved to be highly accurate, since the B parameter is directly related to the ERB spectrum, thus, the bandwidths of the filter correspond to a

fixed distance on the basilar membrane of a human cochlea (Patterson and Moore, 1986). Furthermore, provided the order of the filter is $3 \leq n \leq 5$, the magnitude characteristic of the Gammatone filter closely resembles that of the *roex(p)* filter (Patterson *et al.*, 1992), a standard representation for the magnitude characteristic of the human auditory filter (Patterson and Moore, 1986; Patterson *et al.*, 1988). Typically, while applying the Gammatone filter to represent the frequency sensitivity of a human ear, the order of the filter is set to 4 (Patterson *et al.*, 1987; Lyon *et al.*, 2010). For an even closer approximation, provided the order of the filter n is 4, the bandwidth B should be 1.019 times the ERB at its centre frequency (Patterson and Holdsworth, 1996):

$$B = 1.019 \cdot ERB(f_c) \tag{5.7}$$

Malcolm Slaney's Auditory Toolbox (Slaney, 1993; Slaney, 1998) provided an easy-to-use MATLAB implementation of Patterson's gammatone filter bank, accelerating its use in audio research. The following figure (Figure 5.6) represents the Gammatone filter bank derived using 1024-point Hamming window with an overlap of 512 points, for the frequency range of 0-4000 Hz.

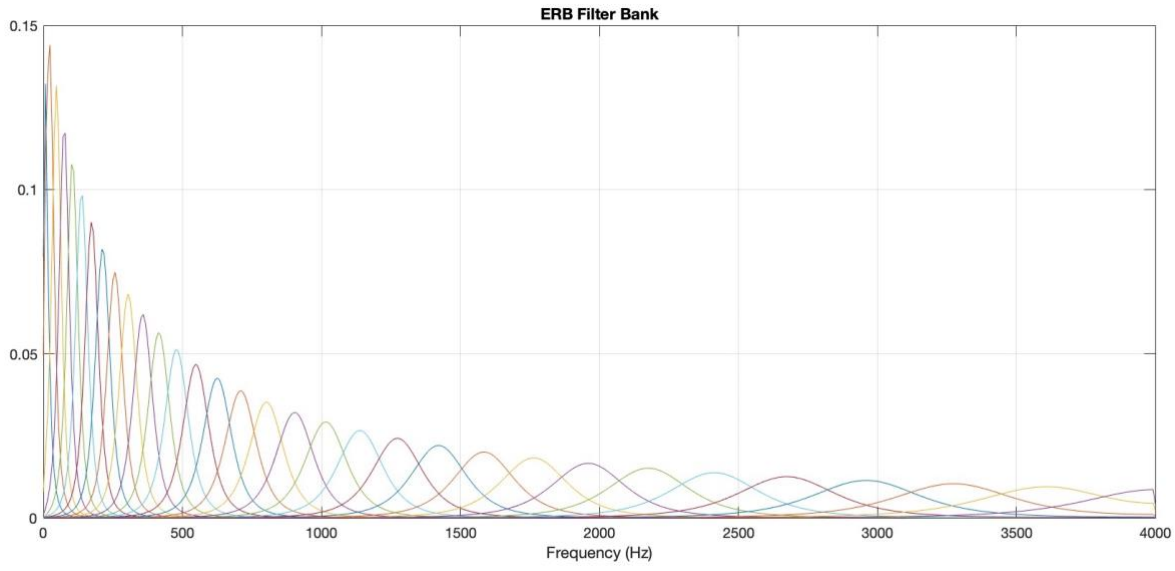

Figure 5.6: Visual representation of a Gammatone (ERB) Filter Bank.

5.3.2 Gammatone Spectrograms ("Gammatonegrams")

A Gammatone spectrogram, also known as a "gammatonegram" or auditory spectrogram (Pour *et al.*, 2014), is a time-frequency representation of a signal obtained using a bank of gammatone filters (Lyon *et al.*, 2010). As an alternative to the conventional spectrogram derived using STFT (where all frequency bins are of equal bandwidths), a Gammatone spectrogram provides a frequency representation that mimics the human ear's varying bandwidths, where their resolution decreases with the increase of the frequency (Ellis, 2009).

Alike in a Mel-spectrogram's case, a Gammatone spectrogram of a signal can be obtained by convolving the signal with a Gammatone filter bank of which the filters are spaced along the ERB scale. The following figure (Figure 5.7) displays a Gammatone spectrogram derived from a continuous speech audio signal obtained from a healthy participant:

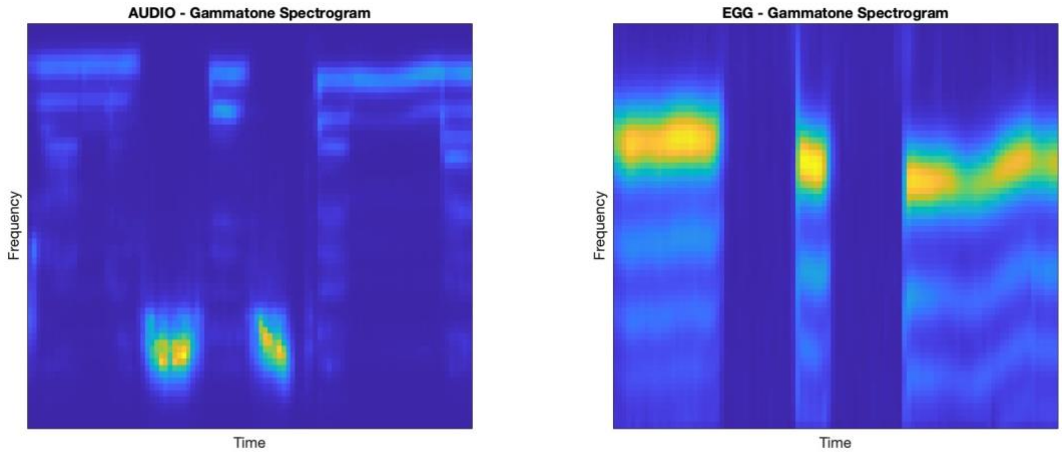

Figure 5.7: Gammatone Spectrograms derived from an audio (left) and bioimpedance signals (right) – a "gammatonegram" derived from a continuous speech signal obtained from a healthy individual.

In this study, the Gammatone spectrograms are used as one of the feature extraction techniques, since they capture the auditory-inspired frequency decomposition more accurately than conventional spectrograms (Ellis, 2009; Lyon *et al.*, 2010; Pour *et al.*, 2014).

All Gammatone spectrograms were derived from the signals subjected to data preprocessing outlined in the previous chapter (including the data segmentation into 1-second-long recordings) with the number of frequency bands equal to 64.

5.3.3 Gammatone Cepstral Coefficients (GTCC)

Gammatone Cepstral Coefficients (GTCCs) are a feature representation and extraction method based on the ERB spectrum, designed to mimic the human auditory system's frequency response, emphasising its nonlinear behaviour. GTCCs can be described as a biologically inspired equivalent of Mel-frequency Cepstral Coefficients, that rely on the application of the Gammatone filter bank, aligned with the ERB frequency bands. Due to their direct relationship with the physiology of human hearing, GTCCs emerged as valuable biology-inspired alternative to MFCCs, traditionally considered the baseline for classification tasks (Bonet-Sola and Alsina-Pages, 2021).

The derivation of GTCCs from a raw signal in time domain follows similar processing to MFCCs derivation. The process consists of following stages: the calculation of the Fourier Transform, application of an appropriate filter bank (Gammatone filter bank for GTCCs), logarithmic compression, and the DCT (Bonet-Sola and Alsina-Pages, 2021).

1. Fourier Transform of the raw signal $x(t)$ in time domain:

The GTCCs derivation process begins with the Fourier Transform of the raw audio signal, capturing its frequency content in different bins.

2. Application of Gammatone filter bank:

Subsequently, the signal undergoes filtering using a Gammatone filter bank, which models the frequency response of the human auditory system. The convolution of the input signal with the impulse response of each Gammatone filter yields filter bank outputs, representing the energy distribution across different frequency bins.

$$y_k(t) = x(t) * g_k(t) \tag{5.8a}$$

$$Y_k(f) = X(f) \cdot G(f) \tag{5.8b}$$

where $X(f)$ represents the frequency spectrum of the signal $x(t)$, $Y_k(f)$ represents the output of the k-th filter of the Gammatone filter bank, and $G(f)$ corresponds to the frequency domain representation of the impulse response of the filter – its frequency response.

The frequency response $G(f)$ of a Gammatone filter can be derived by applying the Fourier Transform directly onto the impulse response $g(t)$ presented above (Equation 5.6). Since the phase shift is usually taken to be zero (Darling, 1991), the equation can be simplified, in which case the frequency response can be represented as follows (Holdsworth *et al.*, 1988):

$$G(f) = \frac{A(n-1)!}{2(2\pi B)^n}\left(\left[1 + \frac{j(f-f_c)}{B}\right]^{-n} + \left[1 + \frac{j(f+f_c)}{B}\right]^{-n}\right) \tag{5.9}$$

where:

- j is the imaginary unit,
- n is the order of the filter,
- B is the bandwidth of the filter,
- f_c is the centre frequency of the filter.

The term $\left[1 + \frac{j(f-f_c)}{B}\right]^{-n}$ represents the contribution of the frequencies below the centre frequency, while the term $\left[1 + \frac{j(f+f_c)}{B}\right]^{-n}$ represents the contribution of the frequencies above the centre frequency.

3. Logarithmic compression (Darling, 1991):

Following this, a unified non-linear compression and logarithmic operation – often referred to as “log-compression” – is applied to simulate the cochlear compression observed in human hearing.

$$Z_k(f) = \log\left(\sum_{k=1}^{K} |Y_k(f)|^2\right) \tag{5.10}$$

where $Z_k(f)$ represents the coefficients of logarithmically compressed energy in k-th filter output, so K is the number of the frequency band in the Gammatone filter bank.

4. Discrete Cosine Transform (DCT):

The log-compressed values are then used as the input for the DCT. The DCT inherently performs dimensionality reduction by transforming the log-compressed filter bank energies into a set of coefficients.

The general DCT mathematical formula most applied in digital signal analysis can be expressed as follows (Ahmed *et al.*, 1974):

$$DCT(i) = \frac{2}{N}\sum_{m=0}^{N-1} x(m) \cdot \cos\left(\frac{(2m+1)i\pi}{2N}\right) \tag{5.11}$$

where:

- $DCT(i)$ is the i-th coefficient of the DCT (i is the index of the DCT coefficient being computed and ranges from 0 to $N-1$),
- N is the total number of data points in the input signal, for $m = 0, 1, 2, \dots, (N-1)$,
- m is the index of the input signal $x(m)$, ranging from 0 to $N-1$,
- $x(m)$ is the input signal. It is a sequence of N data points to be transformed using the DCT.

To obtain the i-th GTCC, the log-compressed filter bank outputs $Z_k(f)$ are used as inputs to the DCT. The formula for the i-th GTCC can therefore be represented as:

$$GTCC(i) = \frac{2}{K}\sum_{k=1}^{K} \alpha_k \, Z_k(f) \cdot \cos\left(\frac{(2k+1)i\pi}{2K}\right) \quad (5.12a)$$

Hence:

$$GTCC(i) = \frac{2}{K}\sum_{k=1}^{K} \alpha_k \log\left(\sum_{k=1}^{K} |Y_k(f)|^2\right) \cdot \cos\left(\frac{(2k+1)i\pi}{2K}\right) \quad (5.12b)$$

Therefore, the final equation of i-th GTCC derivation can be represented within one equation as follows:

$$GTCC(i) = \frac{2}{K}\sum_{k=1}^{K} \alpha_k \log\left(\sum_{k=1}^{K} |X_k(f) \cdot G(f)|^2\right) \cdot \cos\left(\frac{(2k+1)i\pi}{2K}\right) \quad (5.12c)$$

where:

- K represents the number of coefficients in the cosine transform – frequency bins in the Gammatone filter bank,
- i is the index of the specific GTCC coefficient being calculated, which is a feature derived from the input signal using the Gammatone filter bank and further processing,
- k is the index variable of the specific Gammatone filter used in the summation for the cosine transform and it ranges from 1 to N. While i denotes the position of the specific GTCC, k iterates over the number of Gammatone filters, influencing the summation process.
- α_k is the scaling / weighting factor for the k-th Gammatone filter. It is introduced to the equation to provide a way to weight the contribution of each Gammatone filter to the final computation of the GTCCs. This factor helps to emphasise or de-emphasise the importance of individual filters in the feature extraction process.

This equation calculates the i-th GTCC while considering the entire signal as one window. The specific coefficients to be retained are determined by the application

requirements and are often selected based on empirical observations or experimentation with different configurations. This unified process effectively captures the most relevant features from the filter bank outputs, providing a compact representation suitable for further analysis in various audio processing tasks.

GTCCs have gained prominence in speech and audio processing due to their ability to model spectral features in a manner that closely follows the biology of the human auditory system (Lyon *et al.*, 2010; Bonet-Sola and Alsina-Pages, 2021). Due to their close relation to the anatomy of the human cochlea, they are considered particularly effective where auditory-inspired feature extraction is beneficial, such as speech recognition and pathological voice analysis (Valero and Alias, 2012).

In this research, the GTCCs were used as one of the feature extraction methods assessed by the final laryngeal pathology classification system. GTCCs were derived from both data modalities (audio and bioimpedance signals) and preprocessed according to the methodology outlined in the previous chapter (*4.1.3. Data Preprocessing*). The parameters used for the calculation of GTCCs for the purposes of this study are listed in the following table (Table 5.4):

Table 5.4: Parameters defined for GTCC calculation for the developed systems of laryngeal pathology detection and classification.

Data Modality	**Sampling Rate**	**Window Function**	**Window Size**	**Overlap Size**	**Number of Coefficients**
Audio	44100 Hz	Hanning	512	256	40
Laryngeal Bioimpedance	2048 Hz	Hanning	128	64	12

The following figure (Figure 5.8) depicts an example of the GTCC matrices derived from both data modalities for a continuous speech signal obtained from a healthy participant:

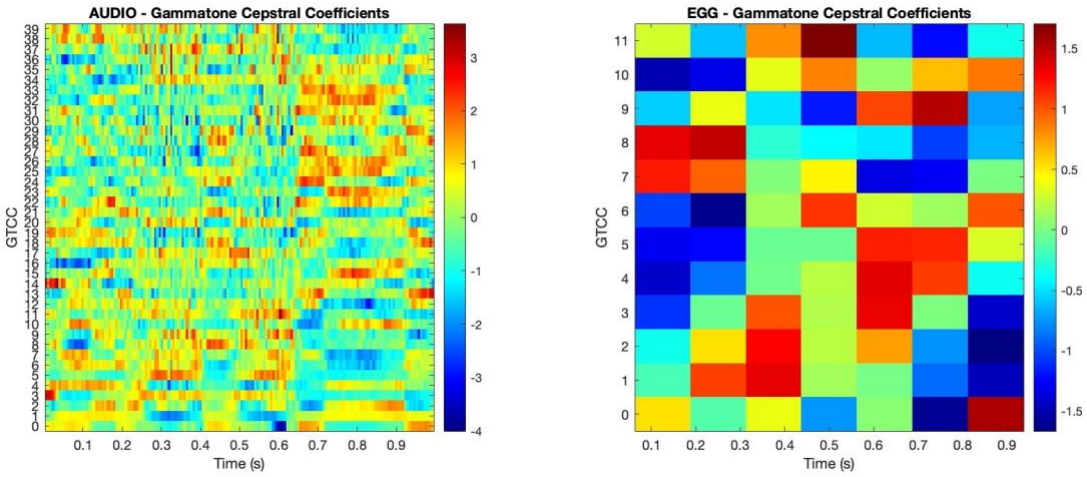

Figure 5.8: Representation of GTCCs derived from an audio (left) and laryngeal bioimpedance (right) recordings of continuous speech obtained from a healthy individual.

Chapter 6

Machine Learning and Deep Learning Methods

The development of a reliable and accurate laryngeal pathology classification system based on audio and laryngeal bioimpedance recordings is the major objective of this research. To fulfil this goal and choose the most appropriate classification methodology, several classification algorithms have been tested and assessed during this study. The following chapter introduces all the investigated methods of digital signal classification in various forms, starting from basic ensemble learners, through to more sophisticated methods relying on deep learning.

First, the Ensemble Learning (EL) methods are explored as the baseline models to assess the feasibility of the derived features (section 6.1). In that section, we introduce a special case of ensemble learning algorithm – the Random Forest (RF) classifier, that delivered the most appropriate approximation of the feature algorithms' performance in the overall classification.

Subsequently, we introduce the deep learning methods – those relying on artificial neural networks (ANN) – used for the development of the laryngeal pathology classifier. We begin with the concept of Convolutional Neural Networks (CNN). Section 6.2 explores three CNN models, each built with a different network architecture. First, the two models built of one-dimensional convolutional layers are introduced (1D-CNN): the "small" CNN model (one with fewer layers), and the "big" CNN model (one with a deeper architecture). Apart from the evaluation of the final dataset, the "big" 1D-CNN model was also used in the preliminary testing of the custom dataset (*4.2. Preliminary Investigation of Custom Dataset Classification* section). Subsequently, the implementation of the CNN model with two-dimensional convolutional layers is introduced (2D-CNN).

Given the sequential nature of human phonation recordings, the implementation of recurrent neural networks (RNN) is also investigated in this chapter. In section 6.3, we introduce the Long Short-Term Memory (LSTM) Network used for the testing of the laryngeal pathology classification, followed by the Bi-directional Long Short-Term Memory (BiLSTM) Network. While LSTM networks generally rely on simpler architecture and require less computation, the BiLSTM enhances the feature learning process by incorporating both past and future context within the signal sequences.

In summary, this chapter provides the description of the methods followed for the development of all classification models tested in terms of performance of laryngeal pathology classification. Furthermore, the parameters used for training and validation processes are discussed in detail.

6. MACHINE LEARNING AND DEEP LEARNING METHODS

6.1. ENSEMBLE LEARNING (EL) AND RANDOM FOREST (RF)

Ensemble learning (EL) is a machine learning technique that combines multiple baseline models, within literature often referred to as “weak learners” or “weak classifiers” (Ferreira and Figueiredo, 2012), to improve the overall predictive performance of a combined model. Generally, a classifier is considered “weak” if it performs only slightly better than a random class assignment (Brownlee, 2021). The fundamental idea is that by aggregating diverse simple models, the ensemble can capture various patterns in the data, leading to more robust and accurate predictions. For instance, if decision trees are used as a base learner, ensembles of decision trees are created, where each is trained on a different subset of data to capture diverse patterns.

A prominent example of an EL algorithm is the Random Forest (RF) classifier, which operates by constructing multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees (Parmar *et al.*, 2019). RF classifiers rely on the technique known as bootstrap aggregating or “bagging”, where each tree in the forest is built using a random subset of the training data, allowing the same data samples to be selected multiple times to form the new random training subsets (Brownlee, 2021). This approach to a training data subset improves generalisation and helps to reduce overfitting, which common decision trees are prone to.

According to related literature, RF demonstrated high accuracy and overall efficacy in various tasks related to distinguishing between vocal characteristics, both in healthy cases (Vaiciukynas *et al.*, 2014), as well as pathological (Wang *et al.*, 2023). Thus, in this study we used RF as a baseline method for evaluating the classification performance based on the feature representations extracted from audio and laryngeal bioimpedance signals (EGG).

The RF classifier was implemented using MATLAB's built-in *TreeBagger* function, with the number of trees ('*NumTrees*') set to 100, and the method set to '*classification*' to enable the categorical outcome prediction.

6.2. CONVOLUTION NEURAL NETWORKS (CNN)

CNNs are a type of ANN designed to imitate the processes of the visual cortex, and as such, they are primarily intended for image recognition tasks (Kim, 2017). The CNNs take their name from the primary operation they perform – convolution, which is defined for two discrete signals $x[n]$ and $h[n]$ with temporal sample index n as:

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] \tag{6.1}$$

In the context of image recognition, the term "convolution" refers to the multiplication process between an input and a filter, represented by a set of weights. A filter, also known as a kernel, is a matrix of specific dimensions that moves across the input data using a specified step size, called "stride". In image processing, the stride typically indicates the number of pixels the filter moves by (Figure 6.1). With each step, the kernel performs the convolution operation on adjacent part of the input data. This results in the production of feature maps that represent the features detected by the kernels. By generating feature maps, the network facilitates feature extraction as each map illustrates the strength and location of the detected property (Patterson and Gibson, 2017).

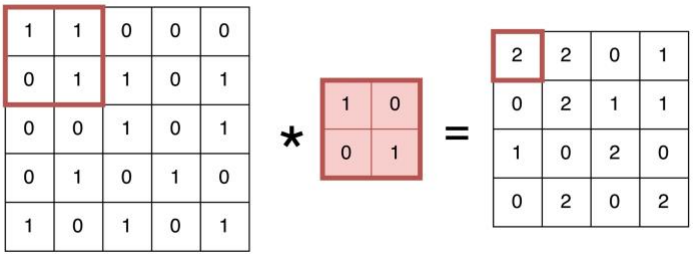

Figure 6.1: Representation of operation of convolutional layer in CNN.

Aside from convolution, CNNs generally rely on the application of pooling layers. Pooling layers are designed to merge adjacent cells of a feature map, thereby reducing the input's dimension. Like convolutional layers, pooling layers also perform a form of convolution. In this case, however, the filter applied is stationary, and the areas affected by the process do not overlap (Figure 6.2).

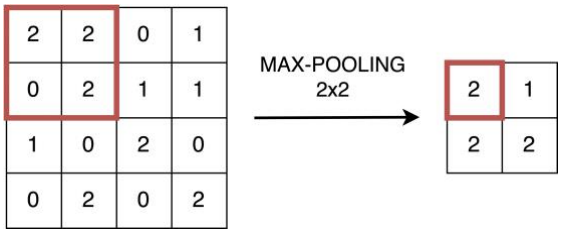

Figure 6.2: Representation of operation of Max-Pooling layer.

There are two main types of pooling: average pooling, which calculates the mean value of the cells within the convolution area, and max-pooling, which identifies the highest value within the convolution area. Additionally, depending on the dimensionality of the data, CNNs may require application of flattening layers, which collapse spatial dimensions of the input into the channel dimension.

While working with CNNs, the Rectified Linear Unit (ReLU) activation function is often used. The function is well-known for preventing the network from the vanishing gradient problem, which occurs when gradients used to update network weights diminish as they are

propagated backward through the network, making it difficult for the network to learn from long sequences. The ReLU function is equal to:

$$f(x) = \max(0, x) \tag{6.2}$$

The ReLU can be defined by a positive argument for itself – it is a non-linear function that outputs the values of the input directly as the output, or zero (Figure 6.3). It is therefore often used as a default activation function, especially while working with CNNs (Kim, 2017).

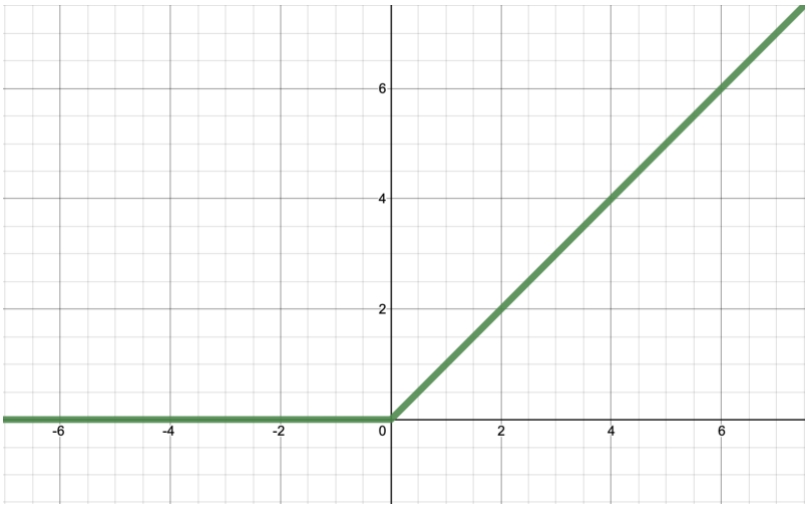

Figure 6.3: Image representation of a Rectifier Linear Unit function.

Although primarily made for image recognition, CNNs became one of most popular classifiers within audio and human-made sounds related research (Aykanat *et al.*, 2017; Muhammad and Alhussain, 2021; Islam *et al.*, 2022; Tomaszewska *et al.*, 2022; Kuo *et al.*, 2023). With the appropriate application of feature extraction algorithms, such as spectrograms or cepstral coefficients (for instance MFCCs and GTCCs), the CNNs can capture both temporal and spectral features, allowing them to discern intricate patterns and representations within sound.

6.2.1 1D-CNN – "Small" Model

The 1D-CNN "small" model is a compact convolutional neural network designed for processing vectorised time-domain and frequency-domain features. Since the model relies on the application of one-dimensional (1D) convolutional layers, the kernels slide along a single dimension of the input making it suitable for both raw sequential data, such as WAV files, and the two-dimensional matrices (for instance cepstral coefficients, or higher order representations such as WVD), reshaped line-by-line into one-dimensional vectors.

This model consists of two blocks of 1D convolutional layers, followed by a normalisation layer, a ReLU activation function, and a dropout layer with the rate of 20% to prevent overfitting. The filters of the convolutional layers applied in this network are, respectively, 20 and 10, with 16 filters per layer. Following the two convolutional blocks, a 1D global average pooling layer is implemented to reduce the temporal dimension, while preserving the most significant activations. The average pooling is followed by a fully connected layer of size $2n$, which stands for twice the initial number of features, a ReLU layer, and another 20% dropout. The network follows into another fully connected layer of size equal to the number of classes (3). The classification is then performed using a softmax layer, followed by the classification layer for the three classes. The following figure (Figure 6.4) represents the architecture of the 1D-CNN "small" model designed for the purposes of this study:

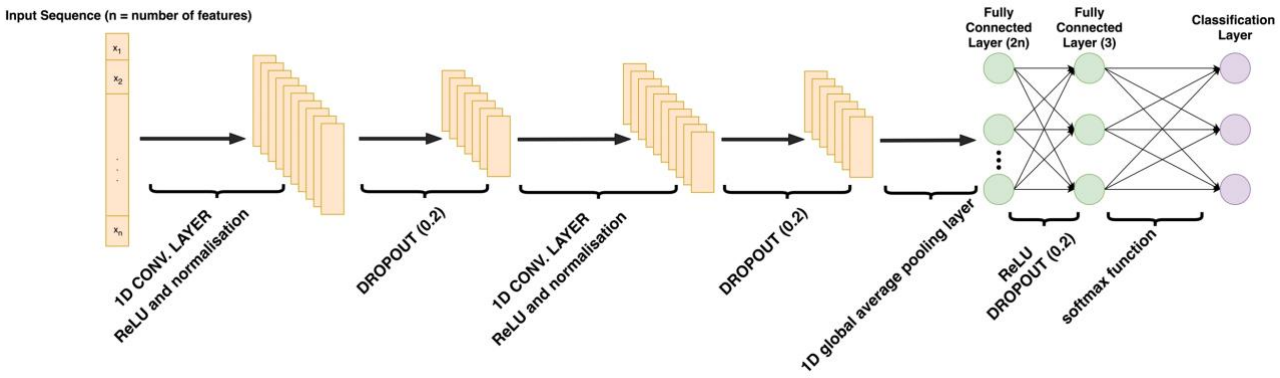

Figure 6.4: The architecture of the 1D-CNN "small" model designed for the purposes of this study and tested as the laryngeal pathology classification system.

As an optimisation algorithm during training, the Adam algorithm was applied. It combines the benefits of two other optimisation methods: AdaGrad and RMSProp, to provide the element-wise moving average of the parameter gradients, as well as their squared values. It therefore uses a similar algorithm to RMSProp, however, with the implementation of the momentum. It also adapts the learning rates for each parameter based on their past gradients and squared gradients. This adaptivity allows for faster convergence and better handling of various optimisations.

The training parameters for the 1D-CNN “small” model were set as follows: mini-batch size equal to 16 samples, training running for the maximum of 50 epochs with shuffling performed at every epoch, and the validation frequency set to 64.

6.2.2 1D-CNN – “Big” Model

The “big” model was built based on a deeper 1D-CNN architecture, featuring increased filter depth and a more extensive convolutional pipeline compared to its “small” counterpart, described in the previous section (1D-CNN “small” model). Importantly, it was used not only for the classification of the final dataset, but also during the preliminary testing of the custom dataset implemented to assess class separability and determine which laryngeal pathology categories must be included in the final system (*4.2. Preliminary Investigation of Custom Dataset Classification* section). Its deeper structure made it suitable for analysing a broader range of acoustic features explored during early-stage data screening.

This model relies on the application of four blocks of 1D convolutional layers with filter sizes of 3 and 5, using 32 and 64 filters in alternating layers to capture both local and broader temporal patterns. Each convolutional layer is followed by the ReLU activation function and the normalisation layer implemented to stabilise the training. The dropout layers with 20% dropout rate are incorporated after the second and the fourth convolutional blocks to mitigate

the overfitting. Subsequently, a 1D global average pooling layer is employed to reduce the sequence dimension prior to fully connected layers. Then, the architecture follows into two fully connected layers with one 50% dropout in between; the first fully connected layer is of size $2n$ equal to twice the initial number of features, and the second is equal to the number of classes – 3. Lastly, the softmax layer is implemented and the data follows into the classification layer.

The following figure (Figure 6.5) depicts the architecture of the 1D-CNN “big” model applied in this study:

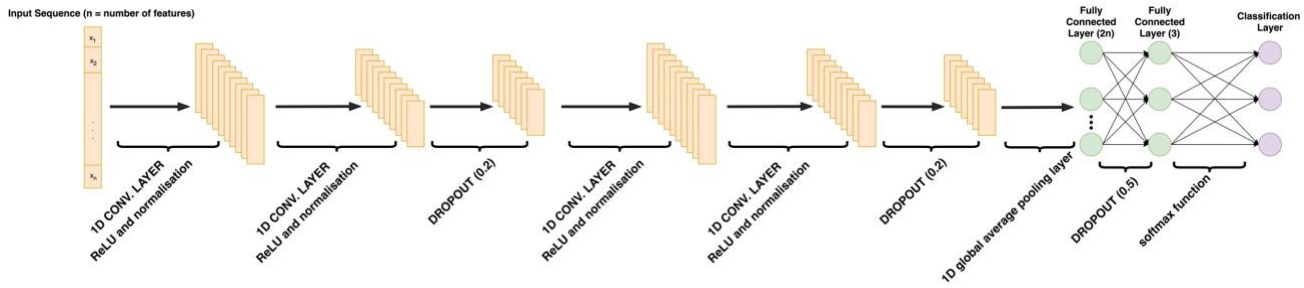

Figure 6.5: The architecture of the 1D-CNN "Big" model designed for the purposes of this study and tested as the laryngeal pathology classification system.

The 1D-CNN “big” model was trained using the Adam optimiser with a mini-batch size of 16 samples, trained for a maximum of 50 epochs, with shuffling employed at every epoch. The validation frequency was set to 64 samples.

6.2.3 2D-CNN Model

The 2D-CNN model was designed for two-dimensional (2D) input representations, such as spectrograms or cepstral coefficients, without the need on feature flattening or concatenation of the matrices. Such representations preserve the spatial information across both time and frequency axes, making a 2D convolutional architecture particularly effective.

The 2D-CNN model developed for the purposes of this study relies on four 2D convolutional blocks, similar in their architecture to those implemented in the 1D-CNN “big”

model; the developed 2D-CNN model relies on four 2D convolutional layers, each followed by a layer normalisation and ReLU activation function, with dropout layers of 20% incorporated in the second and the fourth convolutional blocks to reduce the risk of overfitting. The size of filters in the convolutional layers are 10 (in the first two blocks) and 20 (in the third and the fourth block), while the number of filters alternates between 32 and 64 in every other 2D convolutional layer. Subsequently, a 2D global average pooling is applied to remove the redundant information and aggregate most prominent features into smaller matrices. The model concludes with two fully connected layers – both of size equal to the number of classes – interleaved with a 50% rate dropout layer, before the softmax function and the final classification layer.

The following figure (Figure 6.6) shows the architecture of the implemented 2D-CNN model:

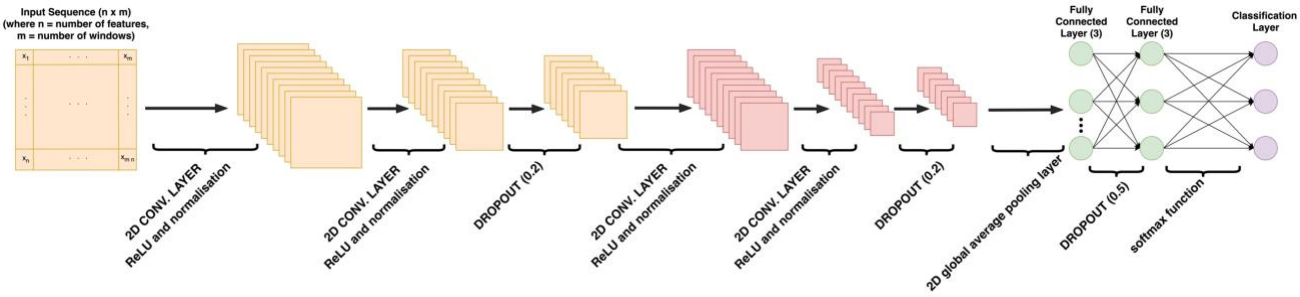

Figure 6.6: The architecture of the 2D-CNN model designed for the purposes of this study and tested as the laryngeal pathology classification system

The 2D-CNN model was trained using the Adam optimisation algorithm, with a mini-batch size of 32, with the maximum epoch number set to 70, with data shuffling applied at every epoch. The validation frequency was set to 64.

6.3. LONG-SHORT TERM MEMORY NETWORKS (LSTM)

LSTM networks are a type of RNN, designed to capture long-range dependencies and sequential information in data. They are particularly well-suited for tasks involving memory,

such as natural language processing, speech recognition, and time series analysis (Patterson and Gibson, 2017). LSTMs had been designed to overcome one of the major issues in traditional RNNs: the vanishing gradient problem. LSTMs solve this problem with gating mechanisms, which allow them to selectively retain and update information from previous time steps (Kim, 2017).

The LSTMs are characterised with the application of memory cells and the gates, including: input gates, which protect nodes from irrelevant inputs; forget gates, which help nodes forget previous memory content; and output gates, which expose memory cell content to the output. Each gate is controlled by sigmoid activation function:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6.3}$$

that produces values between 0 and 1, where 1 means “remember all”, and 0 stands for “forget all”. The memory cells store information over time and can decide what to retain or forget based on the gate activations.

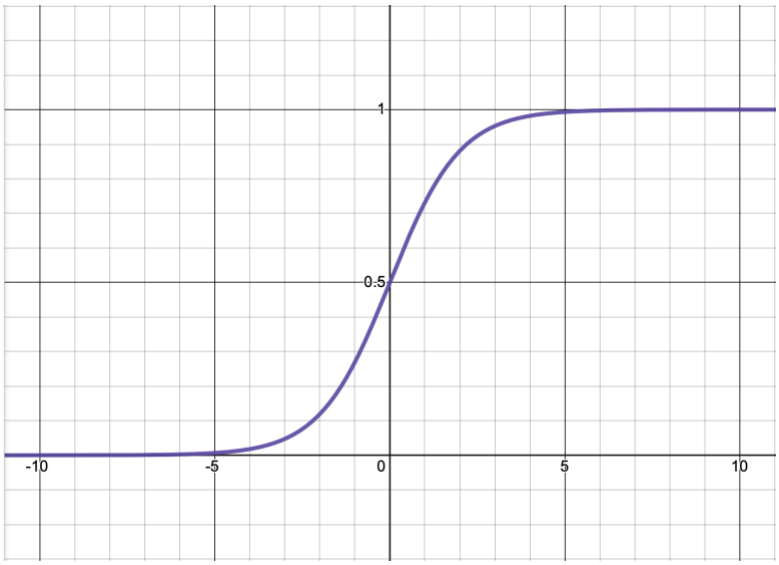

Figure 6.7: Image representation of a Sigmoid activation function.

LSTM networks offer a promising approach for laryngeal pathology classification based on audio recordings due to their unique ability to capture and process sequential information effectively. With their recurrent architecture and memory cells, LSTMs excel at modelling and retaining long-range dependencies in sequential data. This is particularly advantageous in phonatory tract pathology analysis since it allows to capture intricate temporal relationships present in audio signals.

The following figure (Figure 6.8) represents the architecture of the LSTM model designed and implemented for the purposes of this study:

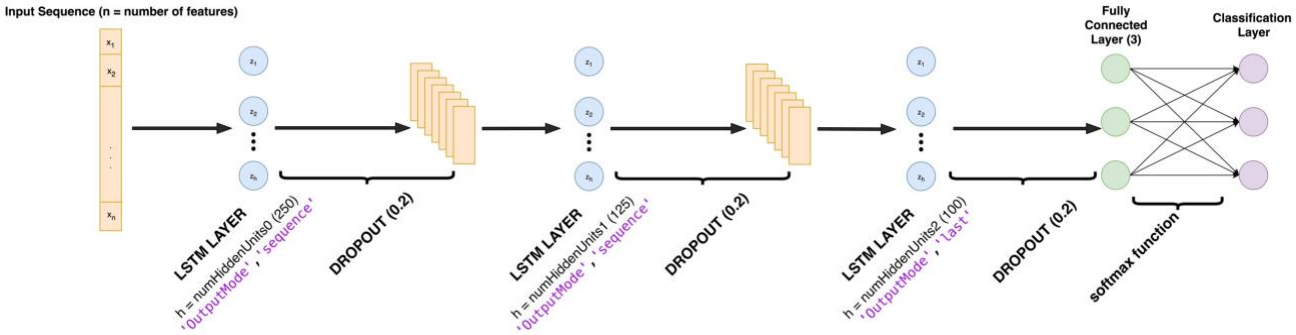

Figure 6.8: The architecture of the LSTM network model designed for the purposes of this study and tested as the laryngeal pathology classification system.

The LSTM model applied in this research relies on three LSTM hidden layers – first of size 250, second 125, and third of size 100 – each followed by a layer of dropout of size 0.2. The choice of incorporating three hidden layers in the LSTM network architecture serves the purpose of capturing complex temporal dependencies within the input data effectively. The inclusion of dropout between these LSTM layers aids in preventing overfitting and ensures the network's generalisation ability.

The training options of the designed LSTM system include the choice of Adam as the optimisation algorithm, as well as the choice of initial learning rate as 0.002 instead of 0.001 to accelerate the training process. Shuffling is performed at every epoch, with the maximum

number of epochs set to 100 to enable the implementation of the early-stopping technique for further prevention of model's overfitting.

6.4. BI-DIRECTIONAL LONG SHORT-TERM MEMORY NETWORKS (BILSTM)

The Bi-directional Long Short-Term Memory (BiLSTM) networks are a type of RNN architecture specifically designed to learn temporal dependencies in both forward and backward directions (Brownlee, 2017). While standard LSTM networks process input sequences in a unidirectional manner, BiLSTMs consist of two parallel LSTM layers: one processes the sequence in the forward order, the other in reverse. This dual processing allows the network to access both past and future contexts for any point in the sequence, leading to improved modelling of temporal patterns.

To exploit the temporal dependencies in both forward and backward directions, a BiLSTM network was also employed in this study. The architecture consists of a sequence input layer followed by a BiLSTM layer with 250 hidden units and a 'last' output mode, enabling the model to capture context from the entire sequence before making a prediction. A dropout layer with a 20% dropout rate follows the recurrent layer to prevent overfitting. The final part of the designed BiLSTM model consists of a fully connected layer that maps the BiLSTM output to the number of classes, followed by a softmax activation and a classification layer. The following figure (Figure 6.9) represents the architecture of the BiLSTM model designed for the purposes of this study:

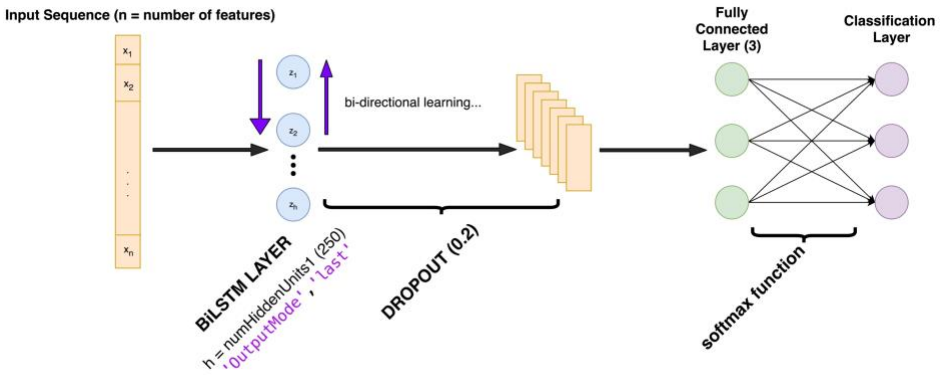

Figure 6.9: The architecture of the BiLSTM network model designed for the purposes of this study and tested as the laryngeal pathology classification system.

This network was trained using the Adam optimiser with the initial learning rate of 0.002 to accelerate the learning process, as was the designed LSTM model described in the previous section. With the shuffling set to occur at every epoch, the maximum epoch number was set to 100 to enable the early-stopping and further mitigate the overfitting. The validation frequency was set to 64 samples.

Chapter 7

Multimodality

Multimodality refers to the integration and combined analysis of multiple heterogeneous data sources, or modalities, such as text, audio, or images, to enhance their processing and the understanding of the information contained within. By combining diverse types of data, multimodal systems aim to capture a more comprehensive representation of the information, leading to improved performance in various tasks, including classification, recognition, or decision-making (Lahat *et al.*, 2015).

In the context of classification tasks, a multimodal approach combines features or parameters derived from multiple data representations – for example audio and laryngeal bioimpedance (EGG) – to improve the recognition of complex patterns and the classifier's decision-making capability. Furthermore, the integration of multimodal data enables better handling of intra-class variability and inter-class overlap. This is particularly relevant in medical diagnostics, where each data modality may be representative of a sole physiological process – for instance, audio signals capture acoustic patterns of an entire phonatory system, while EGG signals provide direct insight into the vocal fold contact dynamics (Mohammed *et al.*, 2023). Therefore, an audio recording obtained from an individual with a laryngeal growth may resemble that of a healthy participant under certain phonation conditions but show marked differences in EGG patterns. The integration of both data modalities provides a more discriminative representation of the underlying class-specific patterns.

At the core of the multimodal machine learning lies the concept of data fusion (Stahlschmidt *et al.*, 2022) – the process of integrating multiple modalities. The appropriate approach to fusing heterogeneous data enables multimodal systems to reinforce consistent

patterns across modalities and compensate for noise or missing information in either source. Thus, the approach to data fusion plays a critical role in shaping the model's performance and interpretability, especially in case of classification systems.

Generally, fusion strategies in multimodal learning can be categorised into three principal approaches: early fusion, hybrid (within literature also referred to as "intermediate" or "middle" – Gadzicki *et al.*, 2020) fusion, and late fusion (Ramachandram and Taylor, 2017). Each approach offers distinct advantages and challenges, depending on the nature of the data and the classification task at hand. In terms of machine learning classification systems, early fusion is when heterogeneous data representations are merged at the level of feature derivation to be processed further as a unimodal input – in this approach, the classification system does not differentiate between modalities, since the individual data are concatenated into a joint representation (Stahlschmidt *et al.*, 2022). In hybrid fusion, the two networks (or alternative classification systems) are merged to provide one output (Gadzicki *et al.*, 2020). In deep learning approaches, this is usually done with the application of flatten and concatenation layers proceeding the fully connected and classification layers. In late fusion, data modalities are processed separately through independent unimodal streams, the results of which are then merged to validate the outputs and provide a higher accuracy of assigned labels (Gadzicki *et al.*, 2020).

In this study, we asses all three fusion methods in the multimodal laryngeal pathology detection and classification. The following chapter discusses the methodology applied during the development of the final multimodal laryngeal pathology classification system. The three fusion strategies are evaluated and compared to choose the most accurate approach to multimodality for laryngeal condition assessment based on combined audio and bioimpedance analysis.

7. MULTIMODALITY

7.1. EARLY FUSION

Early fusion refers to one of the core strategies in multimodal learning where the features of all investigated modalities are merged before feeding into the DL system. The data can be combined raw, for instance as digital signals in WAV format, or in a form of data matrices derived through feature extraction methods, for example as cepstral coefficients. In either approach, the data is concatenated into a single, unified input, that is subsequently processed by a shared classification architecture, seemingly as in a unimodal approach. In early fusion, it is crucial to ensure the appropriate alignment of the data to enable further interpretation of cross-correlation between the modalities (Gadzicki *et al.*, 2020), since the core assumption is that learning cross-modal interactions at a low abstraction level can enable the model to exploit complementary and correlated features across modalities, potentially improving performance in classification or prediction tasks.

Early fusion is especially attractive due to the low complexity of the classification model – the classifier treats the concatenated input as another unimodal representation, with no differentiation between the initial modalities of features (Stahlschmidt *et al.*, 2022). For that reason, no additional branches or feature extraction paths tailored for processing of a specific modality are necessary. This, in turn, allows for a relatively low network complexity by processing all modalities simultaneously. While the low complexity of the system is of benefit, the concatenation of multiple modalities at the level of feature extraction causes the potential for increased dimensionality, making it challenging for models to generalise well, especially in cases of limited data. Furthermore, early fusion has known drawbacks in handling varying temporal resolutions and heterogeneous feature scales (for instance, misalignment in sampling rates or feature dimensionality) – issues often encountered in real-

world biomedical data such as time-series data from audio and electroglottography (Mohammed *et al.*, 2023).

Despite these challenges, when aligned appropriately, early fusion allows for the capture of temporal and spectral patterns across multiple modalities while retaining low complexity and high interpretability. This makes it an advantageous technique for working with digital signals captured simultaneously from various sources, such as simultaneously recorded audio and glottal bioimpedance.

Systematically, a multimodal classification system that relies on an early fusion approach to two data modalities of audio and bioimpedance can be represented as follows (Figure 7.1):

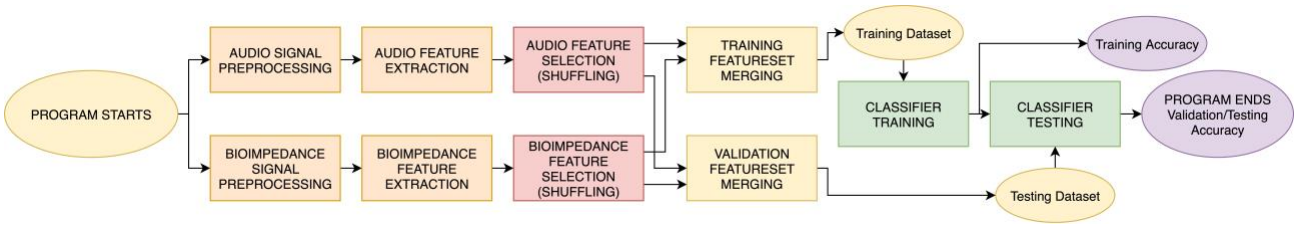

Figure 7.1: Early Fusion Multimodal Classification System developed for the detection and classification of laryngeal pathologies based on combined audio and laryngeal bioimpedance signals.

In this study, the early fusion model was implemented by first extracting the modality-specific features from both audio and EGG signals. Based on results obtained from the unimodal classification systems, further described in detail in chapter 8 of this thesis, we chose GTCCs and Gammatone spectrograms as the feature extraction algorithms for audio and bioimpedance data, respectively. Each signal modality was processed separately using a custom feature extraction pipeline, each described in detail in chapter 5. The derived representations of both modalities were subsequently merged at the feature level via vertical concatenation. This was handled by a custom function *modality_merge_EF* written specifically for the purposes of the early fusion application of this study (the algorithm can be seen in the algorithms folder created for the purposes of this research and is available

upon request). The function zero-pads the shorter feature vector (in terms of time steps) to align dimensionalities before the concatenation. Thus, each fused sample consisted of a vertically stacked matrix of size $N_{audio\ coeffs} + N_{EGG\ coeffs}$ rows and T columns, where T is the number of time steps. The results were stored as a time-series matrix per sample.

The merged features were then inputted into the designed classification system. Based on the results of the unimodal systems, further discussed in chapter 8 of this thesis, the 1D-CNN models were chosen for the testing of the early fusion multimodal approach. The designed classifier was tested first for the detection of the pathological signals, thus, for binary classification between pathological and healthy signals. Subsequently, the classifier was assessed for its multi-class classification performance, distinguishing between cancerous and precancerous growths, neuromuscular diseases, and healthy cases.

The results of the early fusion multimodal classification of laryngeal pathologies based on the above approach are discussed in chapter 9 of this thesis.

7.2. HYBRID (INTERMEDIATE) FUSION

Hybrid fusion, also commonly referred to as intermediate or “middle” fusion (Gadzicki *et al.*, 2020), represents a powerful compromise between early and late fusion strategies in multimodal learning. Unlike early fusion, which focuses on merging various data sources at the feature level, hybrid fusion merges learned modality-specific representations within the body of the model – typically at a mid-depth layer of a deep learning network (Gadzicki *et al.*, 2020). This approach to data fusion allows the network to retain modality-specific learning, thus, the ability to distinguish between various modalities, while also enabling the joint representation of all data. This balances the modality specificity with the cross-modal interaction. The intermediate fusion is particularly suited for applications involving modalities

where signals have inherently different temporal and spectral characteristics and require unique processing pathways (Mohammed *et al.*, 2023). Hybrid fusion has also been favoured while dealing with modalities exhibiting representation alignment changes, such as varying input dimensionality – for instance, WAV file representation for one modality, and spectrograms for another.

Although hybrid fusion appears to have multiple advantages over other fusion methods, its effectiveness and comparison with other strategies is significantly underreported (Lahat *et al.*, 2015). A primary drawback of an intermediate fusion is the increased computational complexity.

According to terminology proposed by Stahlschmidt *et al.* (2022), in the intermediate fusion approach to multimodality the actual data fusion occurs after learning marginal data representations through modality-specific branches – thus, the network branches created specifically for each data modality learn their representations of the data. These are subsequently concatenated (or integrated) to form a joint representation that feeds into shared decision-making layers – fully connected and classification layers. Reportedly, the network fusion is especially beneficial while implemented at the flattened or pooled feature level, enabling the model to exploit both local and global patterns (Gadzicki *et al.*, 2020). The described structure allows for flexibility in processing heterogeneous data types, for instance time-series versus image-like data, using specialised subnetworks that are tailored for each domain.

The flow of a hybrid multimodal learning model that was built for the purposes of this study can be represented with a flow diagram (Figure 7.2):

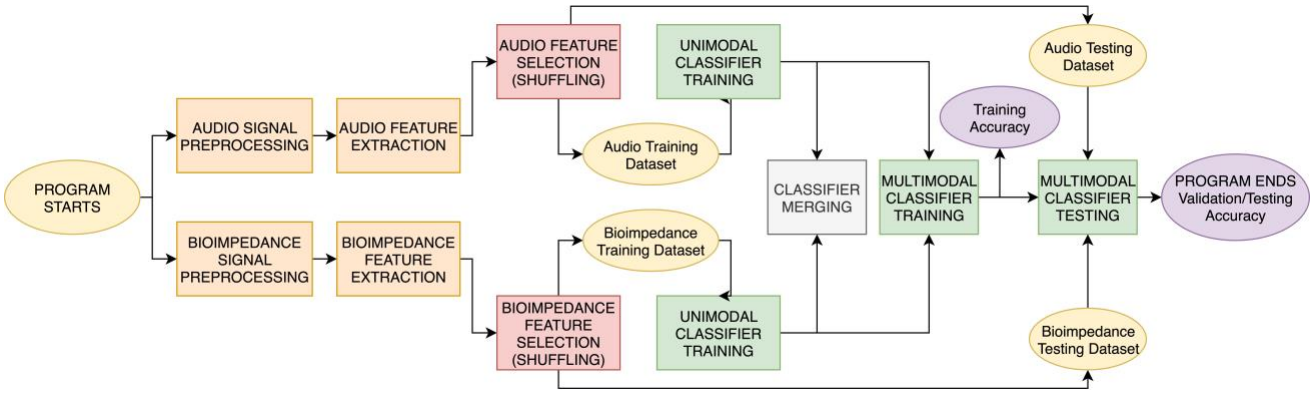

Figure 7.2: Hybrid (Intermediate) Fusion Multimodal Classification System developed for the detection and classification of laryngeal pathologies based on audio and laryngeal bioimpedance data presented in this study.

For the purposes of this study, a hybrid fusion multimodal network was developed for detection and classification of laryngeal pathologies based on combined analysis of audio and laryngeal bioimpedance signals. The intermediate strategy was implemented using parallel DL branches (or "subnetworks") optimised for their respective data modalities, subsequently fused through concatenation layers at an intermediate stage. Both subnetworks were designed based on the unimodal CNN classification systems introduced in chapter 6 of this thesis. Each branch was intended to process its modality independently until the concatenation layer, where both subnetworks implemented flatten layers intended for reshaping the resulting feature maps into vectors most suitable for fusion (Gadzicki *et al.*, 2020). To enable the seamless fusion of the networks, prior to feeding the features into subnetworks, the data was organised into paired training and validation sets, and loaded into array datastore objects, which were then combined for multimodal batching.

For each modality, the architecture of the subnetwork as well as the feature extraction method were chosen based on the results obtained from the unimodal systems, further discussed in chapter 8 of this thesis. The audio signals were subjected to feature extraction in the form of GTCCs, following the methodology outlined in chapter 5. Since the "small" 1D-

CNN delivered the highest accuracy for unimodal classification of the audio signals, the audio subnetwork was designed following its architecture.

Based on the results obtained for bioimpedance signals' classification using the unimodal systems, further described in chapter 8 of this work, the representation of the bioimpedance signals was tested in two forms – Gammatone spectrograms, derived following the methodology outlined in chapter 5, as well as raw WAV files. The architecture of the EGG subnetwork was based on the 2D-CNN model described in chapter 6.

Figure 7.3 depicts the architecture of the laryngeal pathology detection and classification system employing a hybrid fusion multimodal approach, designed for this study.

The modality-specific branches were fused following the pooling layers, aligning with the recommendations of the literature (Gadzicki *et al.*, 2020). At the fusion point, the feature maps were flattened, and the subnetworks were fused using the concatenation layer; the parameters for the concatenation were specified as 1 and 2, signifying the concatenation along the first dimension of two inputs. The network then follows into two fully connected layers, interleaved with a 20% rate dropout and ReLU activation, before the softmax function and the final classification layer for prediction.

The Adam optimisation algorithm was used for training of the hybrid multimodal detection and classification system built for this study. The training was set to 51 epochs, shuffling at every epoch, with the mini batch size of 16 and the validation frequency of 64 samples. The training was pursued using array datastore objects for synchronised audio and EGG inputs. This architecture allowed each modality to retain its own representational learning path, while still enabling the network to exploit cross-modal correlation through the fusion of flattened intermediate features. The use of flatten layers was particularly important in this context, as they ensured that the two branches could be merged coherently into a unified representation vector.

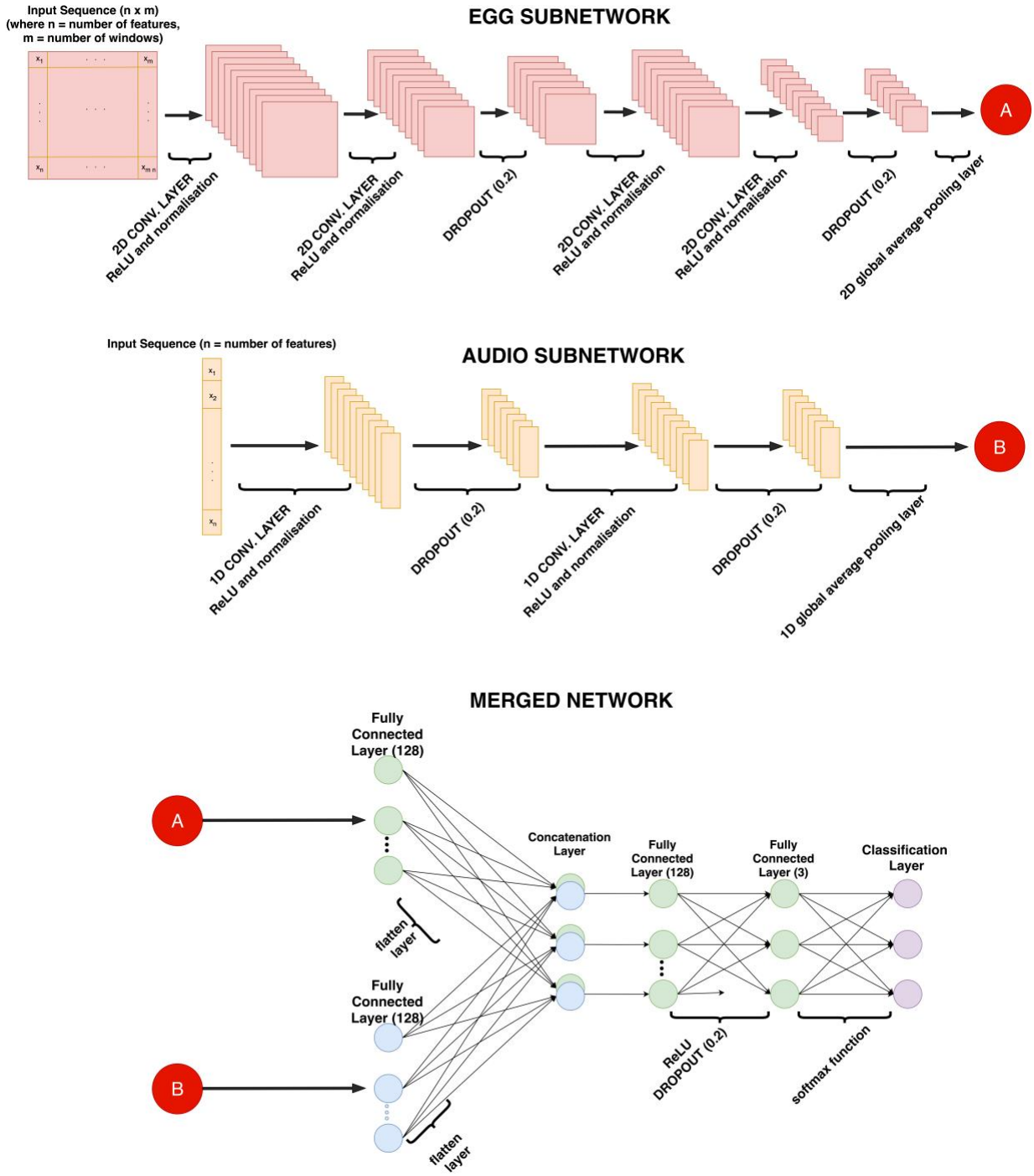

Figure 7.3: The architecture of the Hybrid Multimodal laryngeal pathology detection and classification system designed for this study.

The results of testing the multimodal system implementing the hybrid fusion approach are documented in chapter 9 of this thesis.

7.3. LATE FUSION

Late fusion strategies in multimodal deep learning are characterised by the independent processing of each modality through entirely separate neural networks, followed by the fusion of their respective predictions at the output level. This approach is particularly beneficial when modalities exhibit substantial heterogeneity or have differing temporal and spatial characteristics – for example, in the case of two signal modalities recorded separately (not simultaneously), limiting the possibility of time step alignment. Instead of learning joint or coordinated representations early in the architecture, late fusion maintains a modality-specific learning pipeline. This allows for robust targeted learning approaches tailored to the properties of a given modality. As such, late fusion allows for maximal specialisation of networks to their respective modalities, while still enabling the complete system to learn from multiple data sources.

The main advantages of late fusion strategies include primarily its simplicity and flexibility – the most basic approaches to late fusion include averaging or summation of the modality-specific outputs (Gadzicki *et al.*, 2020). Furthermore, since the fusion takes place at the output of all modalities, there is no need for implementation of high-complexity fusion or concatenation layers requiring increased processing power. Nevertheless, the application of late fusion severely limits the capabilities of a multimodal system to leverage the cross-correlation features found between used modalities. This challenge can be addressed through the stacked generalisation (stacking technique), where an additional classifier is trained at the fusion level with the predictions (often class probabilities) obtained from all investigated unimodal systems (Wolpert, 1992; Stahlschmidt *et al.*, 2022). These approaches are often considered part of decision-level data fusion aimed at increasing certainty and robustness of predictions (Lahat *et al.*, 2015), especially in biomedical and clinical applications.

Late fusion can be categorised into three types depending on the approach taken during the merging of the outputs of the modality-specific models: averaging, weighted fusion, and meta-learning (Stahlschmidt *et al.*, 2022). The averaging, as the name suggests, stands for the process of averaging the probabilities outputted by the softmax function for each class in each modality. In this approach, the contributions of all modalities are equal, since no weighting of the outputs is performed. In the weighted approach, the contribution of each modality towards the final output can be manipulated by adjusting the weights of its predictions. For instance, the predicted probability of each subnetwork can be weighted by its uncertainty, allowing the modalities less prone to errors to have larger impact on the final prediction of the system (Wang *et al.*, 2021).

In the meta-learning approach, an additional "meta-classifier" (a small classification system, often an EL model) is trained on the outputs of independent modality-specific networks (Wolpert, 1992). A meta-classifier is a secondary model that operates on the outputs of the primary classification system, rather than on the raw data input. In the context of late fusion, a meta-classifier learns to interpret the patterns of predictions from each unimodal system – in this case, class probabilities outputted by the softmax function for each class in each modality. The final decision is then made by modelling how these predictions relate to the true class labels. This approach is often referred to as stacked generalisation or stacking (Wolpert, 1992), and it enables the system to resolve conflicts or reinforce agreements between the unimodal classifiers, thus, improving the final system's generalisability and certainty.

The stacked generalisation approach was chosen for the implementation of late fusion for the final multimodal laryngeal pathology detection and classification system investigated in this study. The overall flow of the final multimodal system relying on stacking late fusion that was designed for the purposes of this research can be depicted as follows (Figure 7.4):

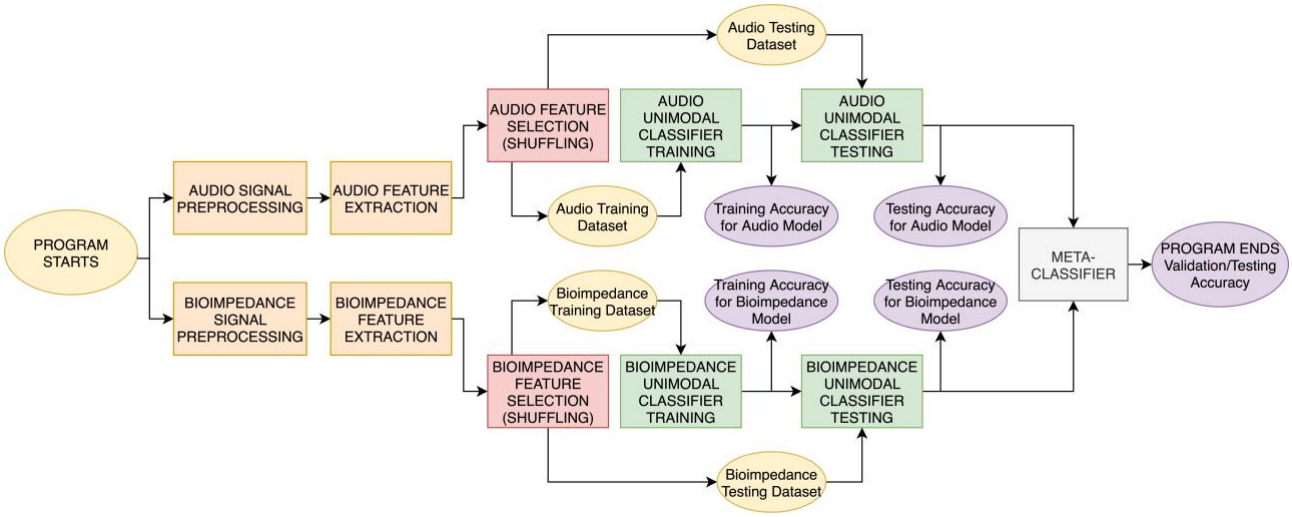

Figure 7.4: Late Fusion Multimodal Classification System developed for the detection and classification of laryngeal pathologies based on audio and laryngeal bioimpedance data presented in this study.

In this study, the stacking late fusion approach to multimodality was implemented using MATLAB, where audio recordings and bioimpedance (EGG) signals were processed through two independent deep learning classification networks. The applied feature extraction methods, as well as the architecture of both modality-specific DL classifiers were based on the methodology described in chapters 5 and 6. Both feature extraction and modality-specific networks were chosen based on the results obtained from the unimodal classification systems' testing performed for each modality – further described in chapter 8 of this thesis. Thus, the audio recordings were processed in a feature form of GTCC matrices, while for the bioimpedance signals two feature representations were examined: the Gammatone spectrograms, as well as the raw WAV file representation. The features extracted from both modalities were then fed into the corresponding DL classification networks, both with the architecture following that of 1D-CNN models. The modality-specific classifiers were trained and validated independently to learn robust unimodal feature representations.

Subsequently, the probabilistic outputs of the networks obtained from the validation datasets – i.e. the outputs of the softmax functions – were extracted and concatenated to

form a feature vector. The newly created vector was then treated as the feature input for the meta-classifier designed to implement the stacking technique, and to generate the final class predictions of the late fusion multimodal classification system.

As the meta-classifier for the designed late fusion multimodal system the error-correcting output codes (ECOC) model was used. ECOC is an ensemble learning method applied in supervised machine learning. It is used primarily for multi-class discrimination problems, where a complex classification task is decomposed into multiple binary classification tasks (Peterson and Weldon, 1972; Dietterich and Bakiri, 1991), using basic binary classification algorithms such as AdaBoost or SVMs (Liu *et al.*, 2015). For the purposes of this research, the ECOC model was designed with the application of $\frac{K(K-1)}{2}$ binary SVMs, where K is the number of unique class labels (*fitcecoc* MATLAB built-in function). Although primarily designed for multi-class problems, since ECOC uses multiple binary classifiers, in this study, the model is used for both detection (binary classification) and multi-class discrimination of laryngeal pathologies.

This approach allowed the designed late fusion multimodal system to maximise the benefits of the presence of two modalities – audio and bioimpedance data. The application of two DL classifiers fused at the stage of classification using stacking enabled robust decision-making even in presence of high uncertainty predictions. The modular design further facilitated the interpretability and flexibility in experimentation, as the underlying unimodal models could be adjusted independently with no additional changes required in the fusion mechanism.

The final accuracy of the stacked classifier was assessed against the ground truth classes. The results obtained using the designed stacked late fusion multimodal system in classification of the laryngeal pathologies is further described and analysed in chapter 9 of this thesis.

Chapter 8

Unimodal System Results

The final goal of this study is the development of a multimodal laryngeal pathology classification system, capable of detecting cancerous and precancerous conditions with high precision and sensitivity, that relies on the combined input of audio recordings of human phonation and the simultaneous recordings of glottal bioimpedance collected from the laryngeal area using electroglottography (EGG).

To compile the most reliable multimodal system for detection and classification of the laryngeal pathologies based on audio and EGG, it is imperative to first develop the unimodal classification systems – relying on each data modality independently – that deliver the best performance possible provided the computational constraints. To build the best-performing unimodal laryngeal pathology classification systems, various approaches to classification models and feature extraction methods were assessed on each data modality independently to choose the most appropriate approach. The following chapter presents the results of the testing completed on various classification models combined with several extracted feature types, following the methodology outlined in chapter 5 and 6. Analysing each modality in isolation allows for a clear assessment of their individual diagnostic potential and establishing the most appropriate classification methodology for each data modality.

Although the final intended outcome of this research is the development of a multi-class laryngeal pathology classification system capable of detecting cancerous and precancerous lesions with high precision and sensitivity, the designed multi-class approach focuses on three limited classes of laryngeal conditions. In case of a pathology detection model, the system must be capable of distinguishing between healthy and any kind of laryngeal disorder, including those omitted in the classification system design. Therefore, the

development of a robust and reliable binary system of laryngeal pathology detection is paramount.

The objective of this study was the development of two robust and accurate systems: one for detection of any pathology present within the laryngeal area (1), and the second for the classification of the laryngeal conditions between cancerous and precancerous, neuromuscular, and healthy states, with the particular focus on the ability to detect the malignant cases (2). This chapter discusses the results obtained for both classification approaches (1 and 2) using each modality independently.

All models of which the performance is investigated further in this chapter were developed and evaluated using several classification frameworks; Random Forest, 1D-CNNs, 2D-CNNs, and RNN – LSTMs and BiLSTMs, all of which are described in chapter 6. All investigated models were implemented in combination with a variety of domain-specific feature extraction techniques, including raw WAV files, STFT spectrograms, GTCCs, MFCCs, Gammatone spectrograms, and Mel-spectrograms, all of which have been detailed in chapter 6.

To ensure the generalisability of the developed models, the designed classifiers were examined using two datasets – the custom dataset developed for the purposes of this study (also referred to as OURs dataset), as well as the Saarbruecken Voice Database (SVD). It is important to note that the testing using SVD was included primarily for completeness and validation of the developed unimodal classification models, rather than as a reliable benchmark. As detailed in Section 4.4.1 (*Limitations of SVD*), this dataset presents several inherent weaknesses, including repetitions of the same subjects and limited representation of certain pathologies – particularly malignant lesions. These limitations undermine its suitability for robust pathology classification, especially in tasks requiring precise differentiation of cancerous or precancerous conditions. Consequently, while SVD results

are presented and discussed for reference, the key methodological decisions and performance assessments are based on findings obtained from the custom dataset, which offers more representative and clinically relevant data.

A key novelty of this research lies in its critical evaluation of the phonation type in laryngeal conditions – specifically, the comparison of the diagnostic potential between continuous speech and sustained vowel phonation. While sustained phonation is commonly used in research settings for its consistency and ease of analysis, one of the primary hypotheses of this work is that speech retains more diagnostically relevant information and offers better discrimination between various pathological states of the larynx. Therefore, this chapter aims not only to evaluate classification performance across the modalities, but also to assess which type of phonation provides a more reliable input for pathology detection and classification. Statistical significance of the speech versus sustained phonation comparison was assessed using one-way analysis of variance (ANOVA), followed by Tukey's honestly significant difference (HSD) post-hoc test, both performed on the results of multi-class laryngeal pathology classification rather than on the binary pathology detection task. By restricting the ANOVA testing to the multi-class case, the results directly demonstrate whether speech signals provide significantly greater diagnostic reliability than sustained phonation in the detection of cancerous and precancerous conditions, which is the overarching objective of this study. To provide fair comparison of phonation type – independently of the classification model to avoid biasing the comparison towards the architecture and feature combination that may favour speech – the best performing speech models and the best performing sustained phonation models were used.

First, this chapter introduces the methods used for the evaluation of the results obtained from all classification models (section 8.1); accuracy, precision, sensitivity, F1 score, and specificity. Subsequently, the chapter is split into two major sections – 8.2 discusses the

pathology detection (binary classification) based on each data modality independently, while 8.3 explores the pathology classification with the focus on malignant cases performed on each modality in separation. Each section is divided into two parts, from which the first (8.2.1 and 8.3.1) pertain to audio recordings as the input modality, while the second (8.2.2 and 8.2.3) assess the laryngeal bioimpedance signals as the input.

The results presented in this chapter are therefore critical in guiding the design of the final multimodal classification system. By identifying the best-performing classifiers, optimal feature sets, and most informative phonation type for each modality, this work establishes the groundwork for a system that is both accurate and clinically meaningful especially in its ability to detect malignant conditions. These findings lay the foundation for the integration of the most promising approaches into the multimodal system described in the next chapter (chapter 9).

8. UNIMODAL SYSTEM RESULTS

8.1. METHODS OF RESULTS ASSESSMENT

In this study, two main approaches to laryngeal signal classification were presented – the pathology detection, which stands for the binary classification between pathological and control signals (control – those gathered from subjects with no laryngeal pathology diagnosis), as well as the multi-class classification based on an underlying laryngeal pathology, with a particular focus on detection of cancerous and precancerous conditions. To enhance the generalisability of the classification models developed for the purposes of this research, all were examined on two different databases (Custom Dataset and SVD). Each database was split into two subsets depending on the phonation type recorded – continuous speech recordings and sustained phonation recordings. The designed classifiers were tested separately depending on the phonation type to examine the potential of speech and sustained phonation in retaining and conveying the pathological features. Both databases comprised audio signals, as well as simultaneous laryngeal bioimpedance recordings. This categorisation resulted in eight independent data subsets that were further examined on all classification systems in both binary and multi-class approaches:

1. Custom dataset of audio speech signals,
2. Custom dataset of audio sustained phonation signals,
3. Custom dataset of EGG speech signals,
4. Custom dataset of EGG sustained phonation signals,
5. SVD of audio speech signals,
6. SVD of audio sustained phonation signals,
7. SVD of EGG speech signals, and
8. SVD of EGG sustained phonation signals.

For each case, the models were tested ten times to perform the 10-fold cross-validation, each time having reshuffled the data for cross-validation purposes. Subsequently, the average confusion matrices were calculated to represent the classification distribution. The classification performance for each of the eight cases was evaluated based on the following parameters:

- Precision (Pr), denoting the accuracy of positive predictions, calculated as the ratio of true positives to the sum of true positives and false positives:

$$Pr = \frac{TP}{TP + FP}$$

- Sensitivity (Sn), measuring the classifier's capability to capture all positive instances, expressed as the ratio of true positives to the sum of true positives and false negatives:

$$Sn = \frac{TP}{TP + FN}$$

- F1-score ($F1$), a measure of balance between precision (Pr) and sensitivity (Sn) represented by their harmonic mean:

$$F1 = \frac{2 \cdot Pr \cdot Sn}{Pr + Sn}$$

- Specificity (Sp), indicating the classifier's ability to correctly identify negative instances, calculated as the ratio of true negatives to the sum of true negatives and false positives:

$$Sp = \frac{TN}{TN + FP}$$

- Accuracy (Acc), indicating the proportion of correctly classified samples among all possible instances:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

All the parameters described above were calculated based on instances of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Considering the multi-class approach, those instances of can be defined as follows:

- TP are the instances correctly predicted as belonging to a specific class. For each class C_i, TP is the number of instances correctly predicted as C_i.
- FP are the instances incorrectly predicted as belonging to a specific class. For each class C_i, FP is the number of instances predicted as C_i while belonging to a different class.
- TN are the instances correctly predicted as not belonging to a specific class. For each class C_i, TN is the number of instances correctly predicted as not C_i.
- FN are the instances incorrectly predicted as not belonging to a specific class. For each class C_i, FN is the number of instances predicted as not C_i, while in fact belonging to C_i.

In addition to reporting the above performance parameters, statistical significance testing was performed to assess whether recordings of continuous speech provide superior diagnostic performance compared to sustained phonation. Given the overarching objective of this work – the detection of cancerous and precancerous conditions – statistical testing was restricted to the multi-class classification task. One-way ANOVA followed by Tukey's HSD post-hoc test were employed to examine whether differences in classification performance between speech and sustained phonation were statistically significant. Results are reported in terms of F-statistics, p-values, confidence intervals (CI), and effect sizes, thereby ensuring that the findings are supported by both descriptive and inferential evidence.

All classification models designed and tested as unimodal laryngeal pathology detection and classification systems based on audio or bioimpedance signals followed the same program flow, depicted in the following figure (Figure 8.1).

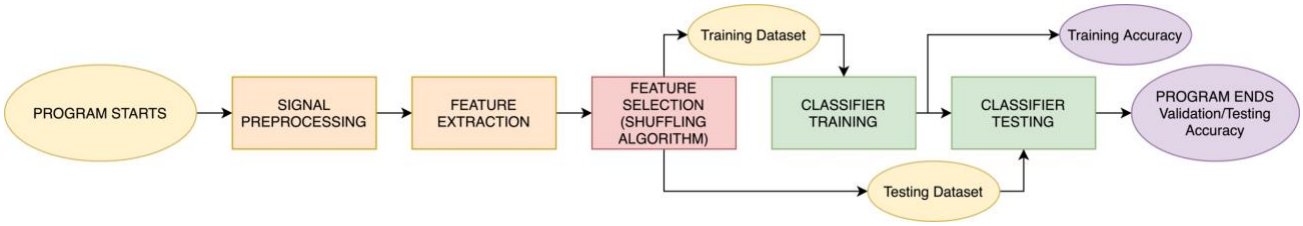

Figure 8.1: Program flow of the audio-based laryngeal pathology detection system.

In summary, two classification approaches were examined: binary and multi-class. Both were examined on two databases: Custom and SVD. Each was split into two data modalities: audio and EGG. Those contained two phonation types: continuous speech and sustained phonation. From each, seven feature forms were extracted: raw WAV files, STFT spectrograms, GTCCs, MFCCs, Gammatone spectrograms, and Mel-spectrograms. These were then tested on the following classification models designed for this study: Random Forest, “small” 1D-CNN, “big” 1D-CNN, 2D-CNN, LSTM, and BiLSTM. In total, the number of classification experiments performed in this study can be expressed as $classification\ type\ (2) \times dataset(2) \times data\ modaity\ (2) \times phonation\ type(2) \times feature\ extraction\ method(6) \times classification\ model(6) = 576$, as follows on the flow diagram below (Figure 8.2).

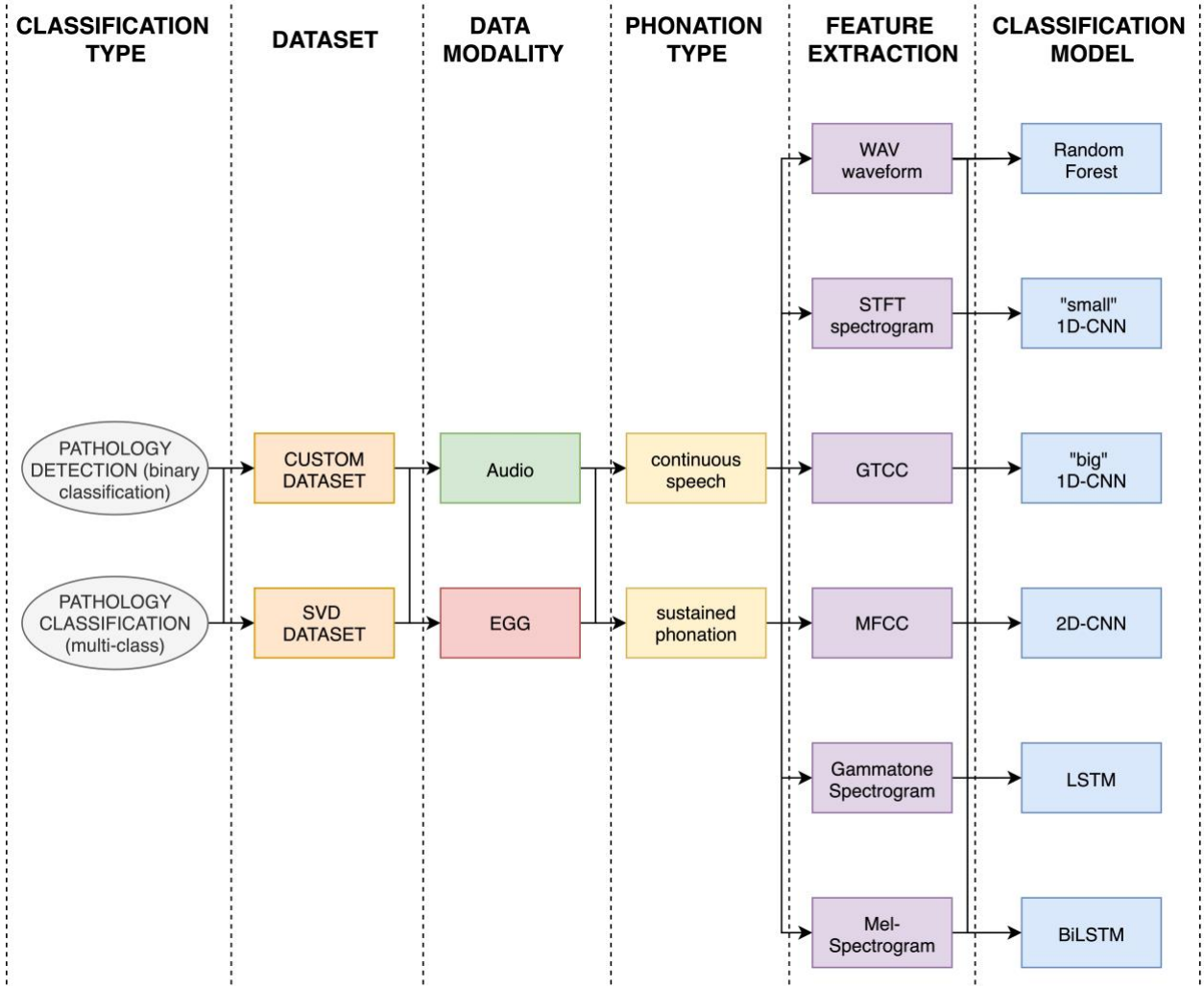

Figure 8.2: The flow of the classification experiments performed in this study for the unimodal detection and classification the laryngeal pathologies.

Due to the high number of classification experiments performed for the purposes of this study, for each type of classification (binary or multi-class) on each data modality (audio or laryngeal bioimpedance) and phonation type (continuous speech or sustained phonation), we first show the overall results in terms of accuracy for each methodology (with respect to feature extraction and classification model), subsequently discussing only the best performing models.

8.2. UNIMODAL LARYNGEAL PATHOLOGY DETECTION

In this section of chapter 8 we discuss the results obtained for the laryngeal pathology detection (binary classification between pathological and control – thus, healthy – signals). The results were obtained from 10-fold cross-validation testing performed on each of the models, according to the methodology outlined in chapters 5 and 6. The performance of the examined classification models fed with different features was examined further, using parameters such as precision, sensitivity, specificity, and F1 score.

8.2.1 Pathology Detection based on Audio Modality

This section is divided according to the classification systems developed and tested as the audio-based laryngeal pathology detection system. In each subsection, the results for all investigated feature extraction algorithms are discussed. Most importantly, for each classification method we present and compare the results obtained from the classification of continuous speech signals and the sustained phonation signals.

8.2.1.1. Random Forest

The results obtained using the RF classifier show moderate accuracy approaching 65%-75% for most of the feature extraction algorithms, with a slight increase in accuracy for the features based on the ERB spectrum (GTCC and Gammatone spectrograms). Notably, the performance of RF drops for MFCC and Mel-spectrograms, which are often considered the standard feature extraction algorithms for speech classification. These results suggest that the ERB spectrum-based features are a more appropriate choice for classification of human phonation signals in medical applications than the Mel spectrum-based feature representations.

While the experiments completed on the custom dataset using RF show better performance of sustained phonation (with maximum of 78.19% ± 5.27 for GTCC), the

continuous speech signals of SVD consistently outperform the sustained phonation (Table 8.1).

Table 8.1: The accuracy of laryngeal pathology detection based on audio, using Random Forest classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	69.71% ± 10.85	67.92% ± 9.60	57.73% ± 6.10	73.41% ± 12.28	56.27% ± 7.32	52.15% ± 1.73
OURs	Sustained Phonation	74.54% ± 5.08	78.19% ± 5.27	71.89% ± 5.29	73.88% ± 4.60	67.59% ± 6.10	59.39% ± 3.76
SVD	Speech	76.35% ± 2.50	77.57% ± 3.24	76.59% ± 2.67	77.87% ± 3.32	77.55% ± 2.69	72.61% ± 2.69
SVD	Sustained Phonation	66.94% ± 3.15	69.22% ± 2.94	67.93% ± 2.05	67.05% ± 2.94	62.34% ± 1.89	60.08% ± 1.53

8.2.1.2. 1D-CNN Classifiers

On average, the laryngeal pathology detection performed using 1D-CNN models outperformed other classifiers designed in this study – seven of the ten highest average accuracy scores were obtained using 1D-CNN. The following table (Table 8.2) presents the results obtained using 1D-CNN, with the seven best instances highlighted in bold and in colour along with the accuracy obtained for the alternative phonation type.

Similar to the results obtained using RF model, the classification performance of 1D-CNNs drops if fed with Mel-spectrum-based features, such as MFCCs or Mel-spectrograms. The highest results were produced for classification of the ERB-spectrum-based feature representations, with GTCCs significantly outperforming others. Notably, in nearly all cases, the accuracy scores were considerably higher for features derived from continuous speech than those derived from sustained phonation. Considering the accuracy of the designed classification models was the highest for 1D-CNN, this finding leads to a conclusion that speech signals significantly outperform the sustained phonation in deep learning-based systems for laryngeal pathology detection.

Interestingly, the raw WAV files produced relatively high accuracy using the “big” 1D-CNN model for both types of phonation; on average, 85.42% ± 5.05 for continuous speech, and 83.77% ± 4.05 for sustained phonation. The classification of raw WAV files did not reach 80% accuracy for any other classification model.

Table 8.2: The accuracy of laryngeal pathology detection based on audio, using 1D-CNN classifiers.

DATASET	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	"small" 1D-CNN	72.74% ± 15.89	**89.17% ± 7.30**	72.02% ± 9.10	76.33% ± 14.85	70.44% ± 15.74	67.49% ± 16.31
OURs	Sustained Phonation	"small" 1D-CNN	77.40% ± 7.42	82.86% ± 4.29	70.45% ± 4.53	76.56% ± 7.41	64.41% ± 12.49	67.21% ± 12.58
OURs	Speech	"big" 1D-CNN	77.97% ± 12.90	**87.33% ± 11.07**	73.15% ± 9.92	78.73% ± 14.17	73.26% ± 11.67	**85.42% ± 5.05**
OURs	Sustained Phonation	"big" 1D-CNN	74.97% ± 8.48	82.79% ± 8.58	72.87% ± 4.46	75.65% ± 6.38	63.48% ± 9.67	**83.77% ± 4.05**
SVD	Speech	"small" 1D-CNN	**80.55% ± 2.32**	**80.67% ± 3.02**	76.79% ± 2.87	**81.30% ± 2.68**	76.12% ± 2.15	56.28% ± 10.15
SVD	Sustained Phonation	"small" 1D-CNN	65.94% ± 2.74	70.79% ± 2.73	68.08% ± 1.87	66.76% ± 3.15	61.13% ± 3.10	55.39% ± 7.00
SVD	Speech	"big" 1D-CNN	76.89% ± 3.35	79.18% ± 3.65	75.32% ± 3.04	78.54% ± 2.69	74.57% ± 3.04	75.95% ± 3.02
SVD	Sustained Phonation	"big" 1D-CNN	66.13% ± 2.72	70.38% ± 3.19	67.98% ± 2.74	66.99% ± 3.32	60.42% ± 2.73	69.16% ± 2.45

For the custom dataset, the highest average accuracy was obtained for GTCCs derived from continuous speech signals and fed into the “small” 1D-CNN model; upon 10-fold cross-validation, the model obtained 89.17% ± 7.30 accuracy, with the best-performing cross-validation instance reaching 97.62%. The table below (Table 8.3) shows the accuracy, precision, sensitivity, specificity, and F1 scores calculated for all instances of the 10-fold cross-validation testing of the best performing model for audio-based laryngeal pathology detection on the custom dataset – the “small” 1D-CNN classification model fed with GTCCs derived from audio speech recordings from the custom dataset:

Table 8.3: The classification parameters calculated for the best performing laryngeal pathology detection system based on audio ("small" 1D-CNN), using GTCCs derived from audio speech data from the custom dataset.

Classification Instance:	Accuracy	Precision	Sensitivity	Specificity	F1
1	94.84%	96.31%	93.25%	96.43%	94.76%
2	87.69%	88.89%	86.15%	89.23%	87.50%
3	89.44%	90.00%	88.73%	90.14%	89.36%
4	82.81%	93.75%	70.31%	95.31%	80.36%
5	75.36%	74.65%	76.81%	73.91%	75.71%
6	97.62%	96.51%	98.81%	96.43%	97.65%
7	91.93%	94.97%	88.54%	95.31%	91.64%
8	96.95%	98.54%	95.31%	98.59%	96.90%
9	93.43%	93.02%	93.90%	92.96%	93.46%
10	81.64%	78.57%	87.01%	76.27%	82.57%
AVERAGE	**89.17%**	**90.52%**	**87.88%**	**90.46%**	**88.99%**
SD	**7.30**	**7.94**	**8.66**	**8.61**	**7.39**

In SVD case, the Gammatone spectrograms derived from continuous speech produced the highest average classification accuracy (81.30% ± 2.68), outperforming the sustained phonation signals (66.76% ± 3.15), achieving the maximum accuracy of 86.15%. The following table (Table 8.4) depicts the accuracy, precision, sensitivity, specificity, and F1 scores calculated for that model using SVD audio speech signals, upon all instances of the 10-fold cross-validation testing:

Table 8.4: The classification parameters calculated for the best performing laryngeal pathology detection system based on audio ("small" 1D-CNN), using Gammatone spectrograms derived from audio speech data from the Saarbruecken Voice Database.

Classification Instance:	Accuracy	Precision	Sensitivity	Specificity	F1
1	77.33%	82.54%	69.33%	85.33%	75.36%
2	82.57%	84.24%	81.76%	83.44%	82.99%
3	81.72%	81.51%	82.07%	81.38%	81.79%
4	86.15%	88.49%	83.11%	89.19%	85.71%
5	82.14%	80.41%	85.00%	79.29%	82.64%
6	80.26%	83.12%	78.53%	82.19%	80.76%
7	80.27%	82.01%	77.55%	82.99%	79.72%
8	77.70%	78.95%	75.54%	79.86%	77.21%

9	80.82%	79.22%	83.56%	78.08%	81.33%
10	84.05%	86.81%	81.17%	87.07%	83.89%
AVERAGE	**81.30%**	**82.73%**	**79.76%**	**82.88%**	**81.14%**
SD	**2.68**	**3.10**	**4.68**	**3.52**	**3.08**

8.2.1.3. 2D-CNN Classifier

Since two-dimensional CNNs excel in learning spatial patterns, the 2D-CNN model designed in this study delivered the highest accuracy for the Gammatone spectrograms, with continuous speech outperforming the sustained phonation (Table 8.5). However, while fed with GTCCs, the 2D-CNN delivered higher accuracy for sustained phonation than the continuous speech signals. Furthermore, relatively high accuracy scores were obtained for raw WAV files.

All results obtained using 2D-CNN fed with audio-derived features were lower than the corresponding accuracy values delivered by 1D-CNN models.

Table 8.5: The accuracy of laryngeal pathology detection based on audio, using 2D-CNN classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	78.10% ± 12.70	79.52% ± 16.49	57.96% ± 11.46	**81.05% ± 12.46**	70.98% ± 12.06	74.28% ± 10.66
OURs	Sustained Phonation	79.83% ± 6.06	**82.30% ± 3.71**	69.69% ± 5.17	80.66% ± 3.49	70.78% ± 8.66	79.55% ± 4.16
SVD	Speech	75.73% ± 3.37	75.35% ± 4.12	68.30% ± 10.31	**80.82% ± 2.48**	77.82% ± 2.74	71.29% ± 3.05
SVD	Sustained Phonation	72.83% ± 2.06	**78.12% ± 3.02**	65.87% ± 5.15	76.94% ± 3.35	71.73% ± 3.10	77.12% ± 2.92

8.2.1.4. RNN Classifiers

The RNN models investigated in this study were LSTM and BiLSTM. The best classification results in the RNN were obtained by the BiLSTM model fed with GTCC matrices derived from continuous speech signals (84.71% ± 11.36 for the custom dataset

and 79.57% ± 2.85 for SVD). The sustained phonation recordings of the custom dataset produced slightly higher accuracy for features including STFT spectrograms, GTCCs, and MFCCs classified using the LSTM model, as well as STFT spectrograms and MFCCs classified using the BiLSTM (Table 8.6). In all remaining testing instances of classification using the RNN continuous speech outperformed sustained phonation.

Table 8.6: The accuracy of laryngeal pathology detection based on audio, using LSTM and BiLSTM classifiers.

DATASET	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	LSTM	72.64% ± 14.83	78.85% ± 12.38	63.61% ± 6.34	75.19% ± 14.39	64.34% ± 11.48	58.96% ± 11.52
OURs	Sustained Phonation	LSTM	73.75% ± 8.05	79.58% ± 5.11	68.88% ± 6.40	71.41% ± 8.68	63.49% ± 5.79	55.52% ± 13.71
OURs	Speech	BiLSTM	73.00% ± 12.26	**84.71% ± 11.36**	69.52% ± 9.48	73.62% ± 12.53	65.45% ± 10.26	64.83% ± 11.46
OURs	Sustained Phonation	BiLSTM	75.15% ± 5.75	79.31% ± 4.59	73.07% ± 4.57	73.53% ± 8.85	65.09% ± 6.50	63.34% ± 12.06
SVD	Speech	LSTM	73.67% ± 2.77	74.42% ± 4.18	73.99% ± 3.32	74.77% ± 3.74	72.22% ± 3.02	56.30% ± 3.48
SVD	Sustained Phonation	LSTM	65.57% ± 2.74	70.04% ± 2.68	66.58% ± 2.13	66.23% ± 3.47	59.60% ± 3.23	53.82% ± 8.87
SVD	Speech	BiLSTM	75.23% ± 4.00	**79.57% ± 2.85**	76.98% ± 2.56	76.63% ± 3.01	73.00% ± 3.57	59.87% ± 3.86
SVD	Sustained Phonation	BiLSTM	65.89% ± 2.64	70.92% ± 3.01	67.11% ± 2.73	67.14% ± 3.13	60.37% ± 2.15	55.67% ± 4.47

Although suited for capturing the sequential dependencies of temporal data, the classification performance of both RNN models designed for the purposes of this study underperform on the classification of the raw WAV files fed as the classifier's input. Furthermore, the classification of audio WAV files of sampling rate 44100 Hz using RNN models is computationally costly and time-consuming. Due to these factors and the lack of promising results for the RNN-based classification of raw audio WAV files, the multi-class assessment of laryngeal pathology classification was not pursued using the designed RNNs.

Although LSTMs and BiLSTMs do show improvement over RF when employing robust features like GTCC or Gammatone spectrograms, they rarely reach the top-tier accuracies

achieved by deeper CNNs. Overall, the designed RNNs appear to underperform when compared to the CNN-based approaches applied in this study.

8.2.1.5. Conclusions on Audio-Based Unimodal Laryngeal Pathology Detection

The following tables (Table 8.7 and Table 8.8) depict the average accuracy (and its SD) of all models designed for laryngeal pathology detection based on audio signals. The results were obtained from 10-fold cross-validation testing performed on each of the models described in chapter 6, fed with each type of features described in chapter 5. For both tables, we highlight the five highest accuracy parameters in bold and in colour along with the average accuracy obtained for the alternative phonation type. The first table (Table 8.7) shows the results obtained from the laryngeal pathology detection based on audio from the custom dataset:

Table 8.7: The accuracy of all models designed for laryngeal pathology detection based on audio modality performed on the custom dataset.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
Audio	Speech	RF	69.71% ± 10.85	67.92% ± 9.60	57.73% ± 6.10	73.41% ± 12.28	56.27% ± 7.32	52.15% ± 1.73
Audio	Sustained Phonation	RF	74.54% ± 5.08	78.19% ± 5.27	71.89% ± 5.29	73.88% ± 4.60	67.59% ± 6.10	59.39% ± 3.76
Audio	Speech	"small" 1D-CNN	72.74% ± 15.89	**89.17% ± 7.30**	72.02% ± 9.10	76.33% ± 14.85	70.44% ± 15.74	67.49% ± 16.31
Audio	Sustained Phonation	"small" 1D-CNN	77.40% ± 7.42	82.86% ± 4.29	70.45% ± 4.53	76.56% ± 7.41	64.41% ± 12.49	67.21% ± 12.58
Audio	Speech	"big" 1D-CNN	77.97% ± 12.90	**87.33% ± 11.07**	73.15% ± 9.92	78.73% ± 14.17	73.26% ± 11.67	**85.42% ± 5.05**
Audio	Sustained Phonation	"big" 1D-CNN	74.97% ± 8.48	82.79% ± 8.58	72.87% ± 4.46	75.65% ± 6.38	63.48% ± 9.67	**83.77% ± 4.05**
Audio	Speech	2D-CNN	78.10% ± 12.70	79.52% ± 16.49	57.96% ± 11.46	81.05% ± 12.46	70.98% ± 12.06	74.28% ± 10.66
Audio	Sustained Phonation	2D-CNN	79.83% ± 6.06	82.30% ± 3.71	69.69% ± 5.17	80.66% ± 3.49	70.78% ± 8.66	79.55% ± 4.16
Audio	Speech	LSTM	72.64% ± 14.83	78.85% ± 12.38	63.61% ± 6.34	75.19% ± 14.39	64.34% ± 11.48	58.96% ± 11.52
Audio	Sustained Phonation	LSTM	73.75% ± 8.05	79.58% ± 5.11	68.88% ± 6.40	71.41% ± 8.68	63.49% ± 5.79	55.52% ± 13.71
Audio	Speech	BiLSTM	73.00% ± 12.26	**84.71% ± 11.36**	69.52% ± 9.48	73.62% ± 12.53	65.45% ± 10.26	64.83% ± 11.46
Audio	Sustained Phonation	BiLSTM	75.15% ± 5.75	79.31% ± 4.59	73.07% ± 4.57	73.53% ± 8.85	65.09% ± 6.50	63.34% ± 12.06

The second table (Table 8.8) presents the results obtained from the laryngeal pathology detection performed on SVD:

Table 8.8: The accuracy of all models designed for laryngeal pathology detection based on audio modality performed on Saarbruecken Voice Database.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
Audio	Speech	RF	76.35% ± 2.50	77.57% ± 3.24	76.59% ± 2.67	77.87% ± 3.32	77.55% ± 2.69	72.61% ± 2.69
Audio	Sustained Phonation	RF	66.94% ± 3.15	69.22% ± 2.94	67.93% ± 2.05	67.05% ± 2.94	62.34% ± 1.89	60.08% ± 1.53
Audio	Speech	"small" 1D-CNN	**80.55% ± 2.32**	**80.67% ± 3.02**	76.79% ± 2.87	**81.30% ± 2.68**	76.12% ± 2.15	56.28% ± 10.15
Audio	Sustained Phonation	"small" 1D-CNN	65.94% ± 2.74	70.79% ± 2.73	68.08% ± 1.87	66.76% ± 3.15	61.13% ± 3.10	55.39% ± 7.00
Audio	Speech	"big" 1D-CNN	76.89% ± 3.35	79.18% ± 3.65	75.32% ± 3.04	78.54% ± 2.69	74.57% ± 3.04	75.95% ± 3.02
Audio	Sustained Phonation	"big" 1D-CNN	66.13% ± 2.72	70.38% ± 3.19	67.98% ± 2.74	66.99% ± 3.32	60.42% ± 2.73	69.16% ± 2.45
Audio	Speech	2D-CNN	75.73% ± 3.37	75.35% ± 4.12	68.30% ± 10.31	**80.82% ± 2.48**	77.82% ± 2.74	71.29% ± 3.05
Audio	Sustained Phonation	2D-CNN	72.83% ± 2.06	78.12% ± 3.02	65.87% ± 5.15	76.94% ± 3.35	71.73% ± 3.10	77.12% ± 2.92
Audio	Speech	LSTM	73.67% ± 2.77	74.42% ± 4.18	73.99% ± 3.32	74.77% ± 3.74	72.22% ± 3.02	56.30% ± 3.48
Audio	Sustained Phonation	LSTM	65.57% ± 2.74	70.04% ± 2.68	66.58% ± 2.13	66.23% ± 3.47	59.60% ± 3.23	53.82% ± 8.87
Audio	Speech	BiLSTM	75.23% ± 4.00	**79.57% ± 2.85**	76.98% ± 2.56	76.63% ± 3.01	73.00% ± 3.57	59.87% ± 3.86
Audio	Sustained Phonation	BiLSTM	65.89% ± 2.64	70.92% ± 3.01	67.11% ± 2.73	67.14% ± 3.13	60.37% ± 2.15	55.67% ± 4.47

The best-performing unimodal laryngeal pathology detection systems based on audio signals were the “small” 1D-CNN models fed with features derived from continuous speech recordings. For the custom dataset, GTCCs outperformed all other feature extraction methods, while SVD delivered the highest accuracy scores for Gammatone spectrograms. The classification parameters, including accuracy, precision, sensitivity, specificity, and F1 scores calculated for the best performing audio-based laryngeal pathology detection systems can be seen in Table 8.3 and Table 8.4.

Based on the results presented in the above sections, the following can be stated – in a laryngeal pathology detection system based on audio recordings of human phonation, on average:

1. The ERB-spectrum-based features such as GTCC and Gammatone spectrograms outperform other feature extraction methods, with those based on the Mel-spectrum delivering significantly lower accuracy scores.
2. Features derived from continuous speech signals outperform those obtained from the recordings of sustained phonation.

This superior performance likely arises from the complex interplay of articulators and broader frequency transitions in speech segments, revealing diagnostic cues that a single sustained vowel cannot match. Among the classifiers, CNN-based methods consistently produce the highest accuracies for speech, with the 1D-CNN models outperforming other methods.

8.2.2 Pathology Detection based on Laryngeal Bioimpedance Modality

This section is divided according to the classification models developed and tested as the EGG-based laryngeal pathology detection systems (binary classification between pathological and healthy laryngeal bioimpedance signals). In each subsection, the results for all investigated feature extraction algorithms are discussed. Most importantly, for each classification method we present and compare the results obtained from the classification of continuous speech signals and the sustained phonation signals.

8.2.2.1. Random Forest

In both datasets, RF yields low to moderate performance for the laryngeal bioimpedance signals, with STFT spectrograms and GTCCs delivering slightly higher results as compared to other feature extraction methods. In SVD, the speech signals provide better accuracy to

sustained phonation, while in the custom dataset sustained phonation outperforms speech. Both phonation type signals of the custom dataset, however, do not exceed 77% accuracy. Furthermore, the highest accuracy score was obtained for GTCCs derived from EGG speech recordings from SVD, reaching 79.11% ± 2.92, outperforming the sustained phonation (Table 8.9). Nevertheless, RF remained outclassed by deep learning models, evaluated in the following sections.

Table 8.9: The accuracy of laryngeal pathology detection based on laryngeal bioimpedance (EGG), using Random Forest classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	61.57% ± 4.45	61.65% ± 6.38	58.14% ± 4.36	61.01% ± 6.27	60.35% ± 6.015	52.88% ± 3.37
OURs	Sustained Phonation	**76.57% ± 8.49**	**76.36% ± 7.28**	71.57% ± 5.47	74.16% ± 7.40	72.33% ± 9.39	57.72% ± 3.92
SVD	Speech	**76.54% ± 3.28**	**79.11% ± 2.92**	75.24% ± 3.07	76.05% ± 2.72	74.67% ± 2.51	70.88% ± 3.00
SVD	Sustained Phonation	73.78% ± 3.48	71.97% ± 3.42	70.54% ± 2.37	71.41% ± 2.68	70.91% ± 2.66	61.19% ± 1.77

8.2.2.2. 1D-CNN Classifiers

The 1D-CNN models provided significantly more accurate results to those obtained using RF. The following table (Table 8.10) presents the results of the EGG-based pathology detection obtained using 1D-CNN, with the instances of the highest average accuracy highlighted in bold and in colour along with the average accuracy obtained for the alternative phonation type.

Notably, the feature extraction methods based on the ERB-spectrum outperform those based on the Mel-spectrum, with Gammatone spectrograms delivering the highest accuracy scores for the custom database, and GTCCs delivering the highest accuracy for SVD. Importantly, in nearly all cases, including all instances of the highest accuracy obtained from the EGG signals tested with all designed classifiers, the features derived from continuous

speech deliver better results to those obtained using sustained phonation. In view of similar results obtained for audio signals, this factor supports the claim that the speech signals provide better representation of the features relevant for the appropriate detection of laryngeal pathologies than those obtained from sustained phonation.

Fed into a deeper 1D-CNN model (the “big” 1D-CNN), raw waveform segments of WAV files also provide robust classification accuracies for both phonation types: for the custom dataset, an average accuracy of 84.02% ± 5.84 was obtained for speech signals, 84.60% ± 4.07 was delivered by sustained phonation, and 78.14% ± 2.40 was reached by the SVD continuous speech data.

Table 8.10: The accuracy of laryngeal pathology detection based on laryngeal bioimpedance (EGG), using 1D-CNN classifiers.

DATASET	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	"small" 1D-CNN	79.75% ± 3.84	83.72% ± 4.54	72.13% ± 5.16	**84.83% ± 6.23**	77.97% ± 3.81	68.71% ± 16.80
OURs	Sustained Phonation	"small" 1D-CNN	80.78% ± 5.03	75.51% ± 8.99	70.81% ± 7.89	82.99% ± 5.65	79.12% ± 6.36	74.87% ± 10.17
OURs	Speech	"big" 1D-CNN	79.69% ± 4.05	81.76% ± 4.27	74.57% ± 7.71	**84.56% ± 4.50**	78.21% ± 4.07	84.02% ± 5.84
OURs	Sustained Phonation	"big" 1D-CNN	81.65% ± 4.96	77.61% ± 5.63	71.45% ± 7.15	82.36% ± 5.87	75.59% ± 9.23	84.60% ± 4.07
SVD	Speech	"small" 1D-CNN	78.40% ± 2.56	**80.99% ± 1.35**	77.78% ± 2.83	**80.55% ± 1.35**	77.02% ± 3.35	63.23% ± 13.34
SVD	Sustained Phonation	"small" 1D-CNN	73.07% ± 3.07	72.59% ± 3.45	70.49% ± 3.31	70.81% ± 2.77	65.89% ± 2.05	65.36% ± 10.95
SVD	Speech	"big" 1D-CNN	78.01% ± 2.99	**80.43% ± 3.03**	77.94% ± 2.54	77.80% ± 3.43	74.91% ± 3.03	78.14% ± 2.40
SVD	Sustained Phonation	"big" 1D-CNN	72.71% ± 2.99	72.86% ± 2.95	71.34% ± 2.86	71.05% ± 2.72	67.86% ± 3.00	73.31% ± 2.82

For SVD, the highest average accuracy scores among all evaluated classifiers were obtained for the “small” 1D-CNN model fed with the GTCC matrices derived from continuous speech. The model delivered an accuracy of 80.99% ± 1.35 on average, outperforming the sustained phonation (72.59% ± 3.45), with the maximum accuracy of 83.22%. The following table (Table 8.11) shows the accuracy, precision, sensitivity, specificity, and F1 scores

calculated for the “small” 1D-CNN model fed with GTCC matrices derived from SVD EGG speech signals, upon all instances of the 10-fold cross-validation testing:

Table 8.11: The classification parameters calculated for the best performing laryngeal pathology detection system based on laryngeal bioimpedance (“small” 1D-CNN), using GTCCs derived from EGG speech data from the Saarbruecken Voice Database

Classification Instance:	Accuracy	Precision	Sensitivity	Specificity	F1
1	79.33%	75.29%	87.33%	71.33%	80.86%
2	80.73%	78.61%	86.47%	74.52%	82.35%
3	79.31%	79.31%	79.31%	79.31%	79.31%
4	83.11%	84.03%	81.76%	84.46%	82.88%
5	80.00%	83.33%	75.00%	85.00%	78.95%
6	81.23%	79.33%	87.12%	74.66%	83.04%
7	80.61%	77.78%	85.71%	75.51%	81.55%
8	80.94%	78.67%	84.89%	76.98%	81.66%
9	83.22%	84.89%	80.82%	85.62%	82.81%
10	81.40%	80.63%	83.77%	78.91%	82.17%
AVERAGE	**80.99%**	**80.19%**	**83.22%**	**78.63%**	**81.56%**
SD	**1.35**	**3.04**	**3.98**	**4.97**	**1.45**

8.2.2.3. 2D-CNN Classifier

The 2D-CNN model achieved some of the highest classification accuracy scores in the laryngeal bioimpedance domain, particularly on the custom dataset. The following table (Table 8.12) presents the results of the EGG-based pathology detection obtained using 2D-CNN, with the instances of the highest average accuracy highlighted in bold and in colour along with the average accuracy obtained for the alternative phonation type.

Notably, the STFT spectrograms derived from the recordings of sustained phonation from the custom dataset delivered the highest accuracy for EGG signal classification (87.39% ± 2.50). Sustained phonation EGG signals from the custom dataset further outperformed continuous speech signals in laryngeal pathology detection as Gammatone spectrograms classified with 2D-CNN (84.88% ± 2.96). Continuous speech EGG signals

obtained from the custom dataset delivered the second highest accuracy in a form of raw waveform segments (WAV files), averaging at 86.03% ± 4.94.

When tested with SVD, the 2D-CNN delivers better accuracy for continuous speech than sustained phonation, with GTCCs outperforming other feature extraction methods (80.98% ± 2.18), followed by STFT spectrograms (79.27% ± 1.56). Nevertheless, the accuracy obtained for SVD with 2D-CNNs did not exceed that of GTCC fed into 1D-CNNs.

Table 8.12 The accuracy of laryngeal pathology detection based on laryngeal bioimpedance (EGG), using 2D-CNN classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	80.82% ± 3.82	80.29% ± 5.11	69.16% ± 5.22	82.14% ± 5.54	77.46% ± 7.53	**86.03% ± 4.94**
OURs	Sustained Phonation	**87.39% ± 2.50**	78.55% ± 4.80	73.33% ± 8.73	**84.88% ± 2.96**	79.36% ± 4.39	82.96% ± 7.84
SVD	Speech	**79.27% ± 1.56**	**80.98% ± 2.18**	77.96% ± 2.06	77.92% ± 2.92	78.81% ± 3.43	73.62% ± 3.03
SVD	Sustained Phonation	74.93% ± 3.27	73.33% ± 1.87	71.27% ± 2.67	78.80% ± 4.07	72.31% ± 3.15	73.92% ± 3.79

The laryngeal bioimpedance signals from the custom dataset performed best when classified using 2D-CNN, with the sustained phonation-derived STFT spectrograms achieving maximum of 92.15% accuracy, and the raw WAVs of speech recordings reaching a maximum of 95.20%. The following tables (Table 8.13 and Table 8.14) present the accuracy, precision, sensitivity, specificity, and F1 scores calculated for all instances of the 10-fold cross-validation testing of this model – the 2D-CNN fed with STFT spectrograms derived from laryngeal bioimpedance recordings of sustained phonation (Table 8.13), followed by the raw WAV files of the continuous speech (Table 8.14):

Table 8.13: The classification parameters calculated for the best performing laryngeal pathology detection system based on laryngeal bioimpedance (2D-CNN), using STFT spectrograms derived from EGG sustained phonation signals from the custom dataset.

Classification Instance:	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**F1**
1	89.27%	89.27%	93.74%	84.17%	94.38%

2	88.44%	88.44%	88.35%	88.56%	88.32%
3	86.43%	86.43%	89.83%	82.17%	90.70%
4	85.03%	85.03%	88.72%	80.27%	89.80%
5	86.03%	86.03%	90.84%	80.13%	91.92%
6	92.15%	92.15%	93.11%	91.03%	93.27%
7	83.20%	83.20%	88.79%	76.00%	90.40%
8	89.12%	89.12%	97.02%	80.72%	97.52%
9	86.97%	86.97%	87.15%	86.72%	87.22%
10	87.29%	87.29%	88.34%	85.91%	88.66%
AVERAGE	**87.39%**	**87.39%**	**90.59%**	**83.57%**	**91.22%**
SD	**2.50**	**2.50**	**3.11**	**4.55**	**3.13**

Table 8.14: The classification parameters calculated for the best performing laryngeal pathology detection system based on laryngeal bioimpedance (2D-CNN), using raw WAV files of EGG speech signals from the custom dataset as input.

Classification Instance:	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**F1**
1	85.71%	87.50%	83.33%	88.10%	85.37%
2	75.90%	82.58%	65.64%	86.15%	73.14%
3	84.74%	91.57%	76.53%	92.96%	83.38%
4	84.90%	92.41%	76.04%	93.75%	83.43%
5	90.40%	97.45%	82.97%	97.83%	89.63%
6	88.10%	97.76%	77.98%	98.21%	86.75%
7	83.85%	97.79%	69.27%	98.44%	81.10%
8	84.74%	90.22%	77.93%	91.55%	83.63%
9	86.72%	94.34%	78.13%	95.31%	85.47%
10	95.20%	98.19%	92.09%	98.31%	95.04%
AVERAGE	**86.03%**	**92.98%**	**77.99%**	**94.06%**	**84.69%**
SD	**4.94**	**5.20**	**7.35**	**4.42**	**5.65**

8.2.2.4. RNN Classifiers

The RNN models developed for this study exhibited moderate to good performance oscillating generally between 70 and 80%, thus below that of CNN-based models. For most instances, continuous speech delivered higher classification accuracy to that of sustained phonation, with the only exceptions being the WAVs and spectrograms (STFT, Gammatone,

and Mel) derived from the custom dataset. The highest accuracy scores were obtained for the GTCCs derived from continuous speech signals for both the custom dataset (81.46% ± 4.61 – LSTM model outperforming BiLSTM), as well as SVD speech (79.09% ± 2.09 – BiLSTM model outperforming LSTM) outperforming the sustained phonation (Table 8.15).

Table 8.15 The accuracy of laryngeal pathology detection based on laryngeal bioimpedance (EGG), using LSTM and BiLSTM classifiers.

DATASET	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	LSTM	76.02% ± 7.95	**81.46% ± 4.61**	70.54% ± 10.13	**78.33% ± 7.04**	71.09% ± 5.85	58.75% ± 9.80
OURs	Sustained Phonation	LSTM	76.83% ± 8.25	74.61% ± 8.29	69.76% ± 9.18	77.98% ± 5.90	73.26% ± 10.03	60.84% ± 11.00
OURs	Speech	BiLSTM	72.76% ± 6.81	79.67% ± 2.98	73.28% ± 5.53	74.92% ± 6.79	66.39% ± 8.04	59.28% ± 5.67
OURs	Sustained Phonation	BiLSTM	77.54% ± 6.77	76.06% ± 3.58	69.26% ± 8.91	78.91% ± 5.32	72.69% ± 6.96	62.48% ± 12.12
SVD	Speech	LSTM	74.87% ± 3.14	78.38% ± 1.84	76.64% ± 2.05	72.75% ± 4.47	67.81% ± 2.43	60.23% ± 5.47
SVD	Sustained Phonation	LSTM	71.24% ± 2.66	72.43% ± 3.08	69.88% ± 2.77	69.39% ± 1.97	65.96% ± 1.95	59.16% ± 5.19
SVD	Speech	BiLSTM	74.16% ± 2.51	**79.09% ± 2.09**	76.74% ± 2.45	**76.85% ± 2.66**	72.73% ± 4.17	65.14% ± 2.79
SVD	Sustained Phonation	BiLSTM	71.23% ± 3.81	72.82% ± 3.75	69.49% ± 2.51	69.00% ± 2.53	65.87% ± 2.18	59.50% ± 6.18

Based on the results above, the RNN models do not perform as well on the laryngeal pathology detection as the proposed CNN-based models. Nevertheless, the results further support the hypothesis of better suitability of the ERB-based features over the features based on the Mel-spectrum in classification of human phonation sounds related to the laryngeal health.

8.2.2.5. Conclusions on Bioimpedance-Based Unimodal Laryngeal Pathology Detection

The following tables (Table 8.16 and Table 8.17) depict the average accuracy (and its SD) of all models designed for laryngeal pathology detection based on laryngeal bioimpedance – EGG signals. The results were obtained from 10-fold cross-validation

testing performed on each of the models, fed with each type of features. For both tables, we highlighted the five highest accuracy scores in bold and in colour along with the accuracy obtained for the alternative phonation type. The first table (Table 8.16) shows the results obtained from the laryngeal pathology detection based on EGG signals from the custom dataset:

Table 8.16: The accuracy of all models designed for laryngeal pathology detection based on laryngeal bioimpedance (EGG) modality performed on the custom dataset.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
EGG	Speech	RF	61.57% ± 4.45	61.65% ± 6.38	58.14% ± 4.36	61.01% ± 6.27	60.35% ± 6.015	52.88% ± 3.37
EGG	Sustained Phonation	RF	76.57% ± 8.49	76.36% ± 7.28	71.57% ± 5.47	74.16% ± 7.40	72.33% ± 9.39	57.72% ± 3.92
EGG	Speech	"small" 1D-CNN	79.75% ± 3.84	83.72% ± 4.54	72.13% ± 5.16	**84.83% ± 6.23**	77.97% ± 3.81	68.71% ± 16.80
EGG	Sustained Phonation	"small" 1D-CNN	80.78% ± 5.03	75.51% ± 8.99	70.81% ± 7.89	82.99% ± 5.65	79.12% ± 6.36	74.87% ± 10.17
EGG	Speech	"big" 1D-CNN	79.69% ± 4.05	81.76% ± 4.27	74.57% ± 7.71	**84.56% ± 4.50**	78.21% ± 4.07	84.02% ± 5.84
EGG	Sustained Phonation	"big" 1D-CNN	81.65% ± 4.96	77.61% ± 5.63	71.45% ± 7.15	82.36% ± 5.87	75.59% ± 9.23	84.60% ± 4.07
EGG	Speech	2D-CNN	80.82% ± 3.82	80.29% ± 5.11	69.16% ± 5.22	82.14% ± 5.54	77.46% ± 7.53	**86.03% ± 4.94**
EGG	Sustained Phonation	2D-CNN	**87.39% ± 2.50**	78.55% ± 4.80	73.33% ± 8.73	**84.88% ± 2.96**	79.36% ± 4.39	82.96% ± 7.84
EGG	Speech	LSTM	76.02% ± 7.95	81.46% ± 4.61	70.54% ± 10.13	78.33% ± 7.04	71.09% ± 5.85	58.75% ± 9.80
EGG	Sustained Phonation	LSTM	76.83% ± 8.25	74.61% ± 8.29	69.76% ± 9.18	77.98% ± 5.90	73.26% ± 10.03	60.84% ± 11.00
EGG	Speech	BiLSTM	72.76% ± 6.81	79.67% ± 2.98	73.28% ± 5.53	74.92% ± 6.79	66.39% ± 8.04	59.28% ± 5.67
EGG	Sustained Phonation	BiLSTM	77.54% ± 6.77	76.06% ± 3.58	69.26% ± 8.91	78.91% ± 5.32	72.69% ± 6.96	62.48% ± 12.12

The second table (Table 8.17) presents the results obtained from the laryngeal pathology detection performed on EGG signals from SVD:

Table 8.17: The accuracy of all models designed for laryngeal pathology detection based on laryngeal bioimpedance (EGG) modality performed on Saarbruecken Voice Database.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
EGG	Speech	RF	76.54% ± 3.28	79.11% ± 2.92	75.24% ± 3.07	76.05% ± 2.72	74.67% ± 2.51	70.88% ± 3.00
EGG	Sustained Phonation	RF	73.78% ± 3.48	71.97% ± 3.42	70.54% ± 2.37	71.41% ± 2.68	70.91% ± 2.66	61.19% ± 1.77

EGG	Speech	"small" 1D-CNN	78.40% ± 2.56	**80.99% ± 1.35**	77.78% ± 2.83	**80.55% ± 1.35**	77.02% ± 3.35	63.23% ± 13.34
EGG	Sustained Phonation	"small" 1D-CNN	73.07% ± 3.07	72.59% ± 3.45	70.49% ± 3.31	70.81% ± 2.77	65.89% ± 2.05	65.36% ± 10.95
EGG	Speech	"big" 1D-CNN	78.01% ± 2.99	**80.43% ± 3.03**	77.94% ± 2.54	77.80% ± 3.43	74.91% ± 3.03	78.14% ± 2.40
EGG	Sustained Phonation	"big" 1D-CNN	72.71% ± 2.99	72.86% ± 2.95	71.34% ± 2.86	71.05% ± 2.72	67.86% ± 3.00	73.31% ± 2.82
EGG	Speech	2D-CNN	**79.27% ± 1.56**	**80.98% ± 2.18**	77.96% ± 2.06	77.92% ± 2.92	78.81% ± 3.43	73.62% ± 3.03
EGG	Sustained Phonation	2D-CNN	74.93% ± 3.27	73.33% ± 1.87	71.27% ± 2.67	78.80% ± 4.07	72.31% ± 3.15	73.92% ± 3.79
EGG	Speech	LSTM	74.87% ± 3.14	78.38% ± 1.84	76.64% ± 2.05	72.75% ± 4.47	67.81% ± 2.43	60.23% ± 5.47
EGG	Sustained Phonation	LSTM	71.24% ± 2.66	72.43% ± 3.08	69.88% ± 2.77	69.39% ± 1.97	65.96% ± 1.95	59.16% ± 5.19
EGG	Speech	BiLSTM	74.16% ± 2.51	79.09% ± 2.09	76.74% ± 2.45	76.85% ± 2.66	72.73% ± 4.17	65.14% ± 2.79
EGG	Sustained Phonation	BiLSTM	71.23% ± 3.81	72.82% ± 3.75	69.49% ± 2.51	69.00% ± 2.53	65.87% ± 2.18	59.50% ± 6.18

The best-performing unimodal laryngeal pathology detection systems based on the laryngeal bioimpedance signals varied for the datasets used. For SVD, the “small” 1D-CNN model fed with GTCCs derived from continuous speech EGG signals outperformed other models, while the custom dataset performed best on 2D-CNN when classified as STFT spectrograms derived from sustained phonation EGG signals, and raw WAVs of EGG continuous speech. The classification parameters, including accuracy, precision, sensitivity, specificity, and F1 scores calculated for the best performing EGG-based laryngeal pathology detection models can be found in Table 8.11 (“small” 1D-CNN on GTCCs from SVD), Table 8.13 (2D-CNN on STFT spectrograms from sustained phonation of custom dataset) and Table 8.14 (2D-CNN on custom dataset’s speech WAVs).

Unlike the clear performance gap observed in audio modality, where speech consistently outperformed the sustained phonation, the results for the EGG data reveal a more balanced dynamic. In several instances, particularly for RF and 2D-CNN models, sustained phonation matches or exceeds the classification performance of speech-derived features. This may be attributed to inherently stable periodic nature of EGG signals, which facilitates the detection and preservation of structural irregularities of the vocal folds behaviour. For the testing

performed on 1D-CNN and RNN classifiers, speech outperforms sustained phonation for the majority of testing instances.

Overall, the CNN-based architectures remain the most effective classifiers for the EGG signals in laryngeal pathology detection. Furthermore, the conclusion on the ERB-spectrum features, such as GTCC and Gammatone spectrograms, outperforming other feature extraction methods, including those based on Mel-spectrum, upholds for both audio and laryngeal bioimpedance signals.

8.3. UNIMODAL LARYNGEAL PATHOLOGY CLASSIFICATION WITH DETECTION OF CANCEROUS AND PRECANCEROUS LESIONS

In this section we discuss the results obtained for the multi-class laryngeal pathology classification system, capable of detecting the cancerous and precancerous lesions. The designed system distinguishes between the three classes: cancerous and precancerous lesions, neuromuscular disorders, and control, thus healthy cases. The section is split further into two subsections, where the first one focuses on the laryngeal pathology classification based on the audio data modality independently, and the second subsection investigates the laryngeal pathology classification using laryngeal bioimpedance as the sole data modality. In each subsection we focus on comparing the classification architecture, feature extraction methods, as well as the phonation type (speech versus sustained phonation) that delivers the highest classification accuracy.

The classification models, described previously in chapter 6, were tested with all feature extraction methods, described previously in chapter 5, with minor exceptions; due to the underwhelming performance of the 2D-CNN model on our custom dataset for multi-class discrimination of laryngeal conditions (further described in the following section), coupled with significant training time and computational requirements of 2D-CNNs, we opted not to conduct testing of the SVD audio dataset using 2D-CNNs. Furthermore, we did not pursue

the classification of the raw WAV files of audio data as the feature input for the designed RNN models. This decision was also made due to the unsatisfactory results of laryngeal pathology detection using audio WAVs and RNNs (as described in results subsection *8.2.1.4. RNN Classifiers* for laryngeal pathology detection based on audio), combined with this method being computationally expensive and time-consuming.

The results presented in this chapter were obtained from 10-fold cross-validation testing performed on each of the unimodal classification systems. The performance of the examined classification models fed with different features was examined using the following parameters: accuracy, precision, sensitivity, specificity, and F1 score. The phonation type comparison is additionally assed using statistical significance testing (for further details, see section *8.3.1.5. Conclusions on Audio-Based Unimodal Laryngeal Pathology Classification*).

8.3.1 Pathology Classification based on Audio Modality

This section is divided according to the discrimination model architectures developed and tested as the audio-based laryngeal pathology classification system capable of detecting cancerous and precancerous lesions. In each subsection, the results for all investigated feature extraction algorithms are discussed. Most importantly, for each classification method we present and compare the results obtained from the classification of continuous speech signals and the sustained phonation signals.

8.3.1.1. Random Forest

Overall, the RF produced low accuracy and inconsistent results when tested on both datasets, with extremely high variances – in some cases, the SD surpassing 96 (Table 8.18). The classifier reached on average between 50 and 60% accuracy, generally with speech signals outperforming those of sustained phonation. Nonetheless, the relatively low accuracy scores and extremely large SD values suggest the RF classifier is not well-suited

for the appropriate laryngeal pathology classification or detection of cancerous and precancerous cases based on human phonation signals, further validating the hypothesis that more sophisticated methods such as deep learning are required for the appropriate classification of the laryngeal pathologies.

Table 8.18: The accuracy of laryngeal pathology classification based on audio, using Random Forest classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	52.48% ± 96.73	63.13% ± 77.80	53.77% ± 49.99	61.72% ± 77.37	48.26% ± 54.69	34.51% ± 21.86
OURs	Sustained Phonation	48.21% ± 80.98	53.38% ± 79.19	48.68% ± 59.01	49.47% ± 70.65	44.86% ± 77.56	42.12% ± 39.50
SVD	Speech	60.05% ± 48.80	55.98% ± 24.67	56.23% ± 22.10	61.80% ± 32.10	57.96% ± 33.32	52.92% ± 16.47
SVD	Sustained Phonation	51.41% ± 24.59	53.41% ± 30.14	52.39% ± 11.88	53.84% ± 32.57	49.63% ± 39.72	39.85% ± 22.24

8.3.1.2. 1D-CNN Classifiers

The 1D-CNN models offered significantly improved performance over RF, delivering 9 out of 10 best average accuracy scores obtained for the unimodal laryngeal pathology classification based on audio signals. The following table (Table 8.19) presents the results, with those 9 instances highlighted in bold and in colour along with the average accuracy obtained for the alternative phonation type.

The acquired results were particularly high for the application of GTCCs, followed by the Gammatone spectrograms, with MFCCs delivering the third best score for SVD audio speech data. For all best (highlighted) classification testing instances performed with 1D-CNNs (and nearly all multi-class discrimination testing instances in total), the performance declined considerably for sustained phonation, generally dropping by more than 10% across both datasets. This trend reflects the models' limited capacity to generalise from sustained static vowel sounds and its reliance on the richer variability present in continuous speech,

leading to a conclusion that speech provides superior discriminatory power in laryngeal condition multi-class discrimination scenarios. This can be attributed to more complex vocal fold dynamics and articulatory movements performed during speech to those performed during sustained vowel phonation.

Table 8.19: The accuracy of laryngeal pathology classification based on audio, using 1D-CNN classifiers.

DATASET	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	"small" 1D-CNN	48.08% ± 8.24	**79.00% ± 7.51**	63.27% ± 7.21	**71.08% ± 6.35**	64.43% ± 6.59	45.68% ± 14.27
OURs	Sustained Phonation	"small" 1D-CNN	48.37% ± 9.25	50.86% ± 6.39	43.98% ± 5.24	54.07% ± 8.64	46.78% ± 11.98	40.09% ± 8.72
OURs	Speech	"big" 1D-CNN	66.14% ± 7.85	**78.83% ± 4.85**	63.25% ± 8.30	**71.35% ± 6.81**	60.05% ± 8.41	63.83% ± 3.56
OURs	Sustained Phonation	"big" 1D-CNN	50.56% ± 11.10	51.93% ± 8.86	42.19% ± 6.97	50.03% ± 10.38	44.29% ± 9.43	59.27% ± 5.16
SVD	Speech	"small" 1D-CNN	66.50% ± 5.98	**71.15% ± 5.68**	**67.59% ± 2.98**	67.51% ± 3.99	58.07% ± 3.53	35.89% ± 8.08
SVD	Sustained Phonation	"small" 1D-CNN	56.86% ± 2.18	61.84% ± 2.74	57.80% ± 2.10	57.19% ± 3.03	48.46% ± 5.12	39.43% ± 7.71
SVD	Speech	"big" 1D-CNN	67.36% ± 5.49	**71.18% ± 5.68**	**67.59% ± 3.87**	**68.01% ± 2.75**	62.37% ± 5.21	61.53% ± 3.99
SVD	Sustained Phonation	"big" 1D-CNN	55.21% ± 2.64	61.26% ± 4.98	58.65% ± 1.32	55.93% ± 3.33	48.39% ± 2.54	53.02% ± 4.68

Based on the results obtained from the custom dataset, the best performing unimodal laryngeal pathology classification system based on audio modality, capable of detecting cancerous and precancerous lesions, is the “small” 1D-CNN model fed with GTCCs derived from continuous speech signals (79.00% ± 7.51). The “big” 1D-CNN fed with speech-derived GTCCs delivered lower SD (78.83% ± 4.85), thus, a more stable performance across all cross-validation instances. However, the obtained accuracy score for this system was slightly lower to that of the “small” 1D-CNN. The maximum accuracy across all cross-validation instances run using 1D-CNNs on the custom dataset was 86.61%, achieved by the “big” 1D-CNN model. The two following tables (Table 8.20 and 8.21) depict the result parameters (accuracy, precision, sensitivity, specificity, and F1 scores) calculated for all instances of the 10-fold cross-validation testing of the two best-performing audio-based

laryngeal pathology classification models tested on the custom dataset – the "small" 1D-CNN classification model (Table 8.20), as well as the "big" 1D-CNN model (Table 8.21). All the result parameters were calculated separately for each class, where class 1 stands for cancerous and precancerous growths, class 2 stands for neuromuscular disorders, and class 3 signifies the healthy cases.

Table 8.20: The classification parameters calculated for one of the two best performing laryngeal pathology classification systems based on audio modality and the custom dataset – the "small" 1D-CNN fed with GTCCs derived from audio speech data. "CA" stands for overall accuracy of the classifier calculated over all three classes. The first column "CI" lists classification instances in cross-validation. The following parameters were numbered according to the class, where 1 stands for cancerous and precancerous, 2 stands for neuromuscular, 3 stand for healthy. "A" is accuracy for the particular class, "P" is precision, "Sn" is sensitivity, "Sp" is specificity, and "F1" is the F1 score.

CI	CA	A1	A2	A3	P1	P2	P3	Sn1	Sn2	Sn3	Sp1	Sp2	Sp3	F1-1	F1-2	F1-3
1	82.46%	83.04%	85.19%	96.69%	68.67%	85.81%	91.46%	76.51%	69.78%	100.00%	85.71%	93.66%	94.86%	72.38%	76.97%	95.54%
2	80.63%	80.82%	82.39%	98.04%	69.47%	73.47%	96.20%	61.07%	79.12%	98.33%	88.95%	84.19%	97.89%	65.00%	76.19%	97.25%
3	81.76%	85.14%	84.46%	93.92%	80.60%	78.01%	85.80%	72.97%	74.32%	97.97%	91.22%	89.53%	91.89%	76.60%	76.12%	91.48%
4	83.63%	83.82%	86.94%	96.49%	78.45%	79.19%	91.00%	61.07%	85.71%	100.00%	93.13%	87.61%	94.56%	68.68%	82.32%	95.29%
5	82.39%	83.56%	84.15%	97.06%	75.59%	75.12%	95.08%	64.43%	82.97%	96.67%	91.44%	84.80%	97.28%	69.57%	78.85%	95.87%
6	70.80%	78.15%	73.25%	90.21%	61.66%	62.00%	90.50%	70.00%	61.69%	80.60%	81.59%	79.51%	95.42%	65.56%	61.85%	85.26%
7	61.70%	65.96%	70.21%	87.23%	49.09%	55.64%	84.80%	57.45%	52.48%	75.18%	70.21%	79.08%	93.26%	52.94%	54.01%	79.70%
8	85.71%	88.12%	88.98%	94.32%	91.78%	79.37%	88.21%	70.16%	90.77%	95.90%	96.92%	88.08%	93.52%	79.53%	84.69%	91.89%
9	76.17%	79.67%	82.24%	90.42%	67.10%	78.81%	83.23%	74.29%	64.58%	89.58%	82.29%	91.20%	90.85%	70.51%	70.99%	86.29%
10	84.73%	87.87%	85.98%	95.61%	77.06%	89.43%	88.65%	87.33%	67.07%	100.00%	88.11%	95.86%	93.31%	81.88%	76.66%	93.98%
AVERAGE	**79.00%**	**81.61%**	**82.38%**	**94.00%**	**71.95%**	**75.69%**	**89.49%**	**69.53%**	**72.85%**	**93.42%**	**86.96%**	**87.35%**	**94.28%**	**70.26%**	**73.87%**	**91.26%**
SD	**7.51**	**6.38**	**6.00**	**3.57**	**11.64**	**10.15**	**4.22**	**8.92**	**11.93**	**8.84**	**7.55**	**5.56**	**2.21**	**8.29**	**9.36**	**5.71**

Table 8.21: The classification parameters calculated for one of the two best performing laryngeal pathology classification systems based on audio modality and the custom dataset – the "big" 1D-CNN fed with GTCCs derived from audio speech data. "CA" stands for overall accuracy of the classifier calculated over all three classes. The first column "CI" lists classification instances in cross-validation. The following parameters were numbered according to the class, where 1 stands for cancerous and precancerous, 2 stands for neuromuscular, 3 stand for healthy. "A" is accuracy for the particular class, "P" is precision, "Sn" is sensitivity, "Sp" is specificity, and "F1" the is F1 score.

CI	CA	A1	A2	A3	P1	P2	P3	Sn1	Sn2	Sn3	Sp1	Sp2	Sp3	F1-1	F1-2	F1-3
1	80.51%	80.90%	83.82%	96.30%	63.08%	87.22%	94.05%	82.55%	63.74%	95.60%	80.22%	94.86%	96.68%	71.51%	73.65%	94.82%
2	81.60%	81.80%	87.67%	93.74%	67.07%	90.48%	87.00%	73.83%	73.08%	96.67%	85.08%	95.74%	92.15%	70.29%	80.85%	91.58%

3	81.31%	83.56%	85.81%	93.24%	77.37%	81.48%	84.30%	71.62%	74.32%	97.97%	89.53%	91.55%	90.88%	74.39%	77.74%	90.63%
4	80.90%	82.46%	81.68%	97.66%	74.79%	69.82%	99.42%	59.73%	85.16%	93.96%	91.76%	79.76%	99.70%	66.42%	76.73%	96.61%
5	81.02%	82.39%	86.50%	93.15%	66.48%	80.87%	98.66%	79.87%	81.32%	81.67%	83.43%	89.36%	99.40%	72.56%	81.10%	89.36%
6	69.41%	73.78%	73.95%	91.08%	53.52%	73.64%	92.13%	89.41%	40.30%	81.59%	67.16%	92.18%	96.23%	66.96%	52.09%	86.54%
7	74.00%	76.83%	76.60%	94.56%	61.75%	72.34%	90.41%	80.14%	48.23%	93.62%	75.18%	90.78%	95.04%	69.75%	57.87%	91.99%
8	77.45%	78.49%	87.95%	88.47%	63.31%	89.31%	86.78%	82.20%	72.82%	77.44%	76.67%	95.60%	94.04%	71.53%	80.23%	81.84%
9	75.47%	82.01%	79.91%	89.02%	76.47%	70.42%	79.04%	65.00%	69.44%	91.67%	90.28%	85.21%	87.68%	70.27%	69.93%	84.89%
10	86.61%	88.91%	86.61%	97.70%	80.12%	85.21%	93.71%	86.00%	73.78%	100.00%	90.24%	93.31%	96.50%	82.96%	79.08%	96.76%
AVERAGE	**78.83%**	**81.11%**	**83.05%**	**93.49%**	**68.40%**	**80.08%**	**90.55%**	**77.03%**	**68.22%**	**91.02%**	**82.95%**	**90.84%**	**94.83%**	**71.66%**	**72.93%**	**90.50%**
SD	**4.85**	**4.09**	**4.88**	**3.26**	**8.49**	**7.98**	**6.39**	**9.38**	**14.03**	**7.88**	**8.09**	**5.03**	**3.76**	**4.63**	**10.16**	**4.97**

For SVD, the highest average accuracy was obtained using GTCCs derived from continuous speech signals and the “big” 1D-CNN model; upon 10-fold cross-validation, the model obtained 71.15% ± 5.68 accuracy, with the highest reaching 78.57%. The table below (Table 8.22) shows the accuracy, precision, sensitivity, specificity, and F1 scores calculated for all instances of the 10-fold cross-validation testing of this model tested with SVD:

Table 8.22: The classification parameters calculated for the best performing laryngeal pathology classification systems based on audio modality and SVD – the “big” 1D-CNN fed with GTCCs derived from audio speech data. “CA” stands for overall accuracy of the classifier calculated over all three classes. The first column “CI” lists classification instances in cross-validation. The following parameters were numbered according to the class, where 1 stands for cancerous and precancerous, 2 stands for neuromuscular, 3 stand for healthy. “A” is accuracy for the particular class, “P” is precision, “Sn” is sensitivity, “Sp” is specificity, and “F1” is the F1 score.

CI	CA	A1	A2	A3	P1	P2	P3	Sn1	Sn2	Sn3	Sp1	Sp2	Sp3	F1-1	F1-2	F1-3
1	77.78%	86.11%	83.33%	86.11%	81.82%	75.00%	76.92%	75.00%	75.00%	83.33%	91.67%	87.50%	87.50%	78.26%	75.00%	80.00%
2	68.89%	80.56%	74.44%	82.78%	69.84%	62.96%	73.02%	73.33%	56.67%	76.67%	84.17%	83.33%	85.83%	71.54%	59.65%	74.80%
3	70.56%	79.44%	81.11%	80.56%	67.69%	72.41%	71.93%	73.33%	70.00%	68.33%	82.50%	86.67%	86.67%	70.40%	71.19%	70.09%
4	65.56%	73.33%	74.44%	83.33%	60.00%	60.00%	80.00%	60.00%	70.00%	66.67%	80.00%	76.67%	91.67%	60.00%	64.62%	72.73%
5	63.73%	77.94%	69.12%	80.39%	65.75%	54.39%	68.92%	70.59%	45.59%	75.00%	81.62%	80.88%	83.09%	68.09%	49.60%	71.83%
6	64.58%	76.39%	70.83%	81.94%	64.00%	57.14%	71.15%	66.67%	50.00%	77.08%	81.25%	81.25%	84.38%	65.31%	53.33%	74.00%
7	69.44%	83.89%	76.67%	78.33%	77.19%	62.16%	71.43%	73.33%	76.67%	58.33%	89.17%	76.67%	88.33%	75.21%	68.66%	64.22%
8	76.39%	87.04%	78.24%	87.50%	86.67%	68.12%	75.86%	72.22%	65.28%	91.67%	94.44%	84.72%	85.42%	78.79%	66.67%	83.02%
9	78.57%	85.12%	86.31%	85.71%	81.63%	76.19%	78.57%	71.43%	85.71%	78.57%	91.96%	86.61%	89.29%	76.19%	80.67%	78.57%
10	76.28%	81.41%	78.21%	92.95%	70.18%	68.75%	90.20%	76.92%	63.46%	88.46%	83.65%	85.58%	95.19%	73.39%	66.00%	89.32%

AVERAGE	**71.18%**	**81.12%**	**77.27%**	**83.96%**	**72.48%**	**65.71%**	**75.80%**	**71.28%**	**65.84%**	**76.41%**	**86.04%**	**82.99%**	**87.74%**	**71.72%**	**65.54%**	**75.86%**
SD	**5.68**	**4.47**	**5.37**	**4.25**	**8.84**	**7.53**	**6.19**	**4.81**	**12.39**	**10.11**	**5.25**	**4.00**	**3.60**	**5.97**	**9.43**	**7.12**

8.3.1.3. 2D-CNN Classifier

The 2D-CNN model delivered moderate results not exceeding 65% while tested with the custom dataset (Table 8.23). The performance was therefore deemed unsatisfactory, providing 1D-CNNs significantly outperforming 2D-CNNs in multi-class approach. The continuous speech signals outperformed the sustained phonation, particularly when fed into the network in the form of ERB-spectrum-based features – Gammatone spectrograms and GTCCs. Additionally, the training of the 2D-CNN model proved to be highly time- and resource-consuming. Due to the poor multi-class discrimination performance of the 2D-CNN on the custom dataset, combined with the high computational demands, the SVD testing using 2D-CNNs was not performed.

Table 8.23: The accuracy of laryngeal pathology classification based on audio, using 2D-CNN classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	49.94% ± 8.43	59.85% ± 13.91	46.41% ± 12.09	64.13% ± 11.85	63.51% ± 6.36	59.39% ± 6.30
OURs	Sustained Phonation	51.82% ± 7.16	48.58% ± 5.95	42.17% ± 8.56	51.69% ± 6.39	51.06% ± 8.59	15.99% ± 25.87

8.3.1.4. RNN Classifiers

While generally outperformed by 1D-CNN models, the RNNs demonstrated competitive performance, particularly on continuous speech data (Table 8.24).

Due to the unsatisfactory results of laryngeal pathology detection using audio WAVs and RNNs (*8.2.1.4. RNN Classifiers*), as well as high computational demands of classification of

raw audio files of 44100 Hz sampling rate, the performance of raw WAV files as the input into the RNNs was not examined.

The highest accuracy obtained using RNNs was achieved by the BiLSTM model fed with GTCCs derived from continuous speech from the custom dataset (74.39% ± 6.15). The GTCCs outperformed other feature extraction methods, with the classification accuracy of speech significantly exceeding that of sustained phonation. Despite the improved results over those achieved by RF and 2D-CNN in laryngeal pathology classification based on audio, the model did not reach 70% when tested with SVD. For both datasets, the 1D-CNN models outperformed the RNNs. The trend of ERB-based features and continuous speech outperforming other methods remained.

Table 8.24: The accuracy of laryngeal pathology classification based on audio, using LSTM and BiLSTM classifiers.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	LSTM	62.29% ± 8.05	70.28% ± 8.16	51.43% ± 8.67	61.90% ± 12.27	52.58% ± 11.30	NaN
OURs	Sustained Phonation	LSTM	43.63% ± 8.53	49.64% ± 5.88	43.82% ± 4.55	49.99% ± 14.15	36.61% ± 4.87	32.12% ± 13.27
OURs	Speech	BiLSTM	62.23% ± 8.40	**74.39% ± 6.15**	61.81% ± 7.22	65.49% ± 7.93	52.88% ± 8.92	NaN
OURs	Sustained Phonation	BiLSTM	48.21% ± 80.98	53.38% ± 79.19	48.68% ± 59.01	49.47% ± 70.65	44.86% ± 77.56	42.12% ± 39.50
SVD	Speech	LSTM	49.72% ± 13.27	64.09% ± 6.91	58.33% ± 6.56	55.64% ± 11.43	43.46% ± 5.69	NaN
SVD	Sustained Phonation	LSTM	53.25% ± 2.58	59.90% ± 2.76	55.56% ± 2.23	53.79% ± 3.96	45.97% ± 2.86	NaN
SVD	Speech	BiLSTM	61.18% ± 8.75	**65.16% ± 3.60**	64.73% ± 3.20	57.72% ± 9.26	43.98% ± 6.23	NaN
SVD	Sustained Phonation	BiLSTM	55.01% ± 2.37	59.57% ± 4.84	56.91% ± 2.49	55.99% ± 2.22	49.06% ± 3.90	NaN

8.3.1.5. Conclusions on Audio-Based Unimodal Laryngeal Pathology Classification

The following tables (Table 8.25 and Table 8.26) depict the accuracy (and its SD) of all models designed for laryngeal pathology classification based on audio signals. For both tables, the five highest average accuracy scores are highlighted in bold and in colour along

with the average accuracy obtained for the alternative phonation signals in the particular methodology. The first table (Table 8.25) shows the results obtained from the laryngeal pathology classification based on audio samples from the custom dataset:

Table 8.25: The accuracy of all models designed for laryngeal pathology classification with detection of cancerous and precancerous lesions based on audio modality performed on the custom dataset.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
Audio	Speech	RF	52.48% ± 96.73	63.13% ± 77.80	53.77% ± 49.99	61.72% ± 77.37	48.26% ± 54.69	34.51% ± 21.86
Audio	Sustained Phonation	RF	48.21% ± 80.98	53.38% ± 79.19	48.68% ± 59.01	49.47% ± 70.65	44.86% ± 77.56	42.12% ± 39.50
Audio	Speech	"small" 1D-CNN	48.08% ± 8.24	**79.00% ± 7.51**	63.27% ± 7.21	**71.08% ± 6.35**	64.43% ± 6.59	45.68% ± 14.27
Audio	Sustained Phonation	"small" 1D-CNN	48.37% ± 9.25	50.86% ± 6.39	43.98% ± 5.24	54.07% ± 8.64	46.78% ± 11.98	40.09% ± 8.72
Audio	Speech	"big" 1D-CNN	66.14% ± 7.85	**78.83% ± 4.85**	63.25% ± 8.30	**71.35% ± 6.81**	60.05% ± 8.41	63.83% ± 3.56
Audio	Sustained Phonation	"big" 1D-CNN	50.56% ± 11.10	51.93% ± 8.86	42.19% ± 6.97	50.03% ± 10.38	44.29% ± 9.43	59.27% ± 5.16
Audio	Speech	2D-CNN	49.94% ± 8.43	59.85% ± 13.91	46.41% ± 12.09	64.13% ± 11.85	63.51% ± 6.36	59.39% ± 6.30
Audio	Sustained Phonation	2D-CNN	51.82% ± 7.16	48.58% ± 5.95	42.17% ± 8.56	51.69% ± 6.39	51.06% ± 8.59	15.99% ± 25.87
Audio	Speech	LSTM	62.29% ± 8.05	70.28% ± 8.16	51.43% ± 8.67	61.90% ± 12.27	52.58% ± 11.30	NaN
Audio	Sustained Phonation	LSTM	43.63% ± 8.53	49.64% ± 5.88	43.82% ± 4.55	49.99% ± 14.15	36.61% ± 4.87	NaN
Audio	Speech	BiLSTM	62.23% ± 8.40	**74.39% ± 6.15**	61.81% ± 7.22	65.49% ± 7.93	52.88% ± 8.92	NaN
Audio	Sustained Phonation	BiLSTM	48.21% ± 80.98	53.38% ± 79.19	48.68% ± 59.01	49.47% ± 70.65	44.86% ± 77.56	NaN

The training of the 2D-CNN model proved to be highly time-consuming and computationally expensive. Given the poor results on the custom dataset and the excessive training and validation times, as well as the high computational demands, we decided not to proceed with testing the SVD dataset using 2D-CNNs. The second table (Table 8.26) presents the results obtained from the laryngeal pathology classification performed on SVD:

Table 8.26: The accuracy of all models designed for laryngeal pathology classification with detection of cancerous and precancerous lesions based on audio modality performed on Saarbruecken Voice Database.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
Audio	Speech	RF	60.05% ± 48.80	55.98% ± 24.67	56.23% ± 22.10	61.80% ± 32.10	57.96% ± 33.32	52.92% ± 16.47
Audio	Sustained Phonation	RF	51.41% ± 24.59	53.41% ± 30.14	52.39% ± 11.88	53.84% ± 32.57	49.63% ± 39.72	39.85% ± 22.24
Audio	Speech	"small" 1D-CNN	66.50% ± 5.98	**71.15% ± 5.68**	**67.59% ± 2.98**	67.51% ± 3.99	58.07% ± 3.53	35.89% ± 8.08
Audio	Sustained Phonation	"small" 1D-CNN	56.86% ± 2.18	61.84% ± 2.74	57.80% ± 2.10	57.19% ± 3.03	48.46% ± 5.12	39.43% ± 7.71
Audio	Speech	"big" 1D-CNN	67.36% ± 5.49	**71.18% ± 5.68**	**67.59% ± 3.87**	**68.01% ± 2.75**	62.37% ± 5.21	61.53% ± 3.99
Audio	Sustained Phonation	"big" 1D-CNN	55.21% ± 2.64	61.26% ± 4.98	58.65% ± 1.32	55.93% ± 3.33	48.39% ± 2.54	53.02% ± 4.68
Audio	Speech	LSTM	49.72% ± 13.27	64.09% ± 6.91	58.33% ± 6.56	55.64% ± 11.43	43.46% ± 5.69	NaN
Audio	Sustained Phonation	LSTM	53.25% ± 2.58	59.90% ± 2.76	55.56% ± 2.23	53.79% ± 3.96	45.97% ± 2.86	NaN
Audio	Speech	BiLSTM	61.18% ± 8.75	65.16% ± 3.60	64.73% ± 3.20	57.72% ± 9.26	43.98% ± 6.23	NaN
Audio	Sustained Phonation	BiLSTM	55.01% ± 2.37	59.57% ± 4.84	56.91% ± 2.49	55.99% ± 2.22	49.06% ± 3.90	NaN

In the multi-class approach to laryngeal pathology classification based on audio modality, the “big” 1D-CNN model produced the highest average accuracy results for SVD, while the “small” 1D-CNN delivered the highest average accuracy for the custom dataset. Consistently, speech signals significantly outperformed sustained phonation, particularly when fed into the classifier in the form of ERB-based features – the GTCCs.

To prove that continuous speech provides significantly better results in laryngeal pathology classification using audio recordings (with particular focus on the multi-class setting with detection of cancerous and precancerous lesions), we performed statistical significance testing using ANOVA and Tukey’s HSD of the best performing speech models and the best performing sustained phonation models were used. Those were:

1. For the custom dataset: speech-derived GTCC processed with “small” 1D CNN (accuracy of 79.00% ± 7.51) and raw WAV files of sustained phonation processed with “big” 1D CNN (accuracy of 59.27% ± 5.16),

2. For SVD: speech-derived GTCC processed with “big” 1D CNN (accuracy of 71.18% ± 5.68) and sustained phonation-derived GTCC processed with “small” 1D CNN (accuracy of 61.84% ± 2.74).

For audio signals from the custom dataset, speech achieved significantly higher classification accuracy than sustained phonation ($F(1,18) = 46.88$, $p = 0.000002$), with a mean difference of 0.197 and 95% CI $= [0.137, 0.258]$. The effect size was large ($\eta_p^2 = 0.723$), indicating that approximately 72% of the variance in the dependent variable can be explained by group membership. The same trend was observed on the SVD dataset, where speech again outperformed sustained phonation ($F(1,18) = 21.88$, $p = 0.00019$), with a mean difference of 0.093 and 95% CI $= [0.051, 0.135]$. Also here the effect size was substantial ($\eta_p^2 = 0.549$), demonstrating a robust advantage of speech-based inputs.

This upholds the conclusion that ERB-based feature extraction algorithms and continuous speech signals are more suitable for laryngeal pathology detection and classification tasks than sustained phonation and Mel-spectrum-based features.

In addition to overall classification accuracy, precision and sensitivity metrics were evaluated for the detection of cancerous and precancerous lesions, given the clinical importance of minimising both false positives and false negatives in this category. On average, the “small” 1D-CNN evaluated on the custom dataset achieved the highest precision at 71.95% ± 11.64, indicating strong reliability in correctly identifying malignant samples without misclassifying non-cancerous cases. However, its sensitivity was slightly lower at 69.53% ± 8.92, suggesting a modest rate of missed detections. In contrast, the “big” 1D-CNN on the same dataset showed a more balanced profile, with lower precision (68.40% ± 8.49) but higher sensitivity (77.03% ± 9.38), indicating improved cancerous case detection rates at the cost of an increased false positive rate. Notably, the “big” 1D-CNN also performed well on the SVD data, achieving a precision of 72.48% ± 8.84 and a sensitivity of

71.28% ± 4.81, reflecting strong generalisation capability and consistent performance across datasets.

All classification parameters (accuracy, precision, sensitivity, specificity, and F1 scores) calculated for the best performing audio-based laryngeal pathology classification models can be found in Table 8.20 (“small” 1D-CNN on GTCCs from custom dataset), Table 8.21 (“big” 1D-CNN on GTCCs from custom dataset) and Table 8.22 (“big” 1D-CNN on GTCCs from SVD).

Overall, these results suggest that the “big” 1D-CNN combined with continuous speech offers the most effective trade-off between precision and sensitivity, particularly for clinical scenarios where maximising true positive rates is critical for early detection of malignancies. Therefore, the CNN-based architectures fed with speech remain the most effective classifiers for the audio signals in laryngeal pathology classification.

8.3.2 Pathology Classification based on Laryngeal Bioimpedance Modality

Alike the previous results sections, this part of the thesis is divided according to the classification systems developed and tested as the EGG-based laryngeal pathology classification system capable of detecting cancerous and precancerous lesions. In each subsection, the results for all investigated feature extraction algorithms are discussed. For each classification method, we also present and compare the results obtained from the classification of continuous speech signals and the sustained phonation signals.

8.3.2.1. Random Forest

The RF demonstrated limited effectiveness for multi-class classification using the EGG data (Table 8.27). Nevertheless, the tendency in speech-derived features outperforming sustained phonation remained, with ERB-spectrum-based features outperforming others (with the highest accuracy obtained for Gammatone spectrograms derived from speech,

achieving an average accuracy of 60.94% ± 37.74 on custom data, and 59.10% ± 51.07 on SVD). The large values of SD confirmed more sophisticated methods of classification, such as deep learning, are required for the appropriate multi-class discrimination of laryngeal conditions.

Table 8.27: The accuracy of laryngeal pathology classification based on laryngeal bioimpedance (EGG), using Random Forest classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	58.24% ± 19.56	60.21% ± 50.32	56.97% ± 24.52	60.94% ± 37.74	59.37% ± 53.42	43.96% ± 24.44
OURs	Sustained Phonation	55.05% ± 45.08	51.50% ± 56.49	48.63% ± 65.37	54.27% ± 44.95	52.46% ± 44.80	44.97% ± 60.06
SVD	Speech	54.46% ± 33.09	58.07% ± 43.43	54.44% ± 20.82	59.10% ± 51.07	55.76% ± 31.05	50.65% ± 26.23
SVD	Sustained Phonation	50.54% ± 25.42	53.16% ± 34.84	47.59% ± 31.34	52.26% ± 35.43	49.56% ± 28.55	42.97% ± 21.94

8.3.2.2. 1D-CNN Classifiers

The 1D-CNN models significantly improved the classification results for the multi-class discrimination between the three chosen laryngeal conditions, producing 9 out of 10 best average accuracy scores achieved for the EGG signals. The following table (Table 8.28) presents the results, with those 9 instances highlighted in bold and in colour along with the average accuracy obtained for the alternative phonation type.

The “small” 1D-CNN delivered slightly better results to those obtained with the “big” 1D-CNN model, with continuous speech consistently outperforming the sustained phonation for all classification iterations throughout the cross-validation process. The ERB-spectrum-derived features produced the best results on both datasets, with Gammatone spectrograms achieving the highest scores on the custom dataset (74.21% ± 5.41 average accuracy), and GTCCs outperforming other features while tested using SVD (66.81% ± 6.14).

Table 8.28: The accuracy of laryngeal pathology classification based on laryngeal bioimpedance (EGG), using 1D-CNN classifiers.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	"small" 1D-CNN	**71.88% ± 5.03**	**72.90% ± 5.00**	71.88% ± 6.25	**74.21% ± 5.41**	67.05% ± 4.86	66.18% ± 16.91
OURs	Sustained Phonation	"small" 1D-CNN	56.19% ± 6.92	48.52% ± 7.45	43.68% ± 6.54	52.94% ± 5.49	53.06% ± 7.04	50.66% ± 12.77
OURs	Speech	"big" 1D-CNN	66.13% ± 2.96	**72.05% ± 4.43**	68.41% ± 4.01	69.70% ± 3.83	67.33% ± 3.09	**73.73% ± 5.01**
OURs	Sustained Phonation	"big" 1D-CNN	54.89% ± 12.03	48.59% ± 6.15	48.11% ± 9.91	54.12% ± 9.64	54.85% ± 8.44	57.17% ± 8.42
SVD	Speech	"small" 1D-CNN	57.42% ± 7.74	**66.81% ± 6.14**	59.13% ± 6.06	**63.21% ± 5.93**	58.35% ± 4.48	40.52% ± 9.07
SVD	Sustained Phonation	"small" 1D-CNN	50.65% ± 3.63	56.36% ± 2.21	49.46% ± 3.50	53.29% ± 3.27	52.00% ± 3.01	39.75% ± 9.17
SVD	Speech	"big" 1D-CNN	59.10% ± 4.78	**62.27% ± 4.28**	57.99% ± 7.46	61.35% ± 5.38	57.72% ± 5.26	**64.04% ± 6.08**
SVD	Sustained Phonation	"big" 1D-CNN	50.23% ± 3.80	55.28% ± 4.01	50.10% ± 3.98	53.38% ± 3.24	50.43% ± 4.01	56.19% ± 3.23

During testing of the custom dataset, the best performance for unimodal laryngeal condition classification based on laryngeal bioimpedance was obtained for speech-derived Gammatone spectrograms and the "small" 1D-CNN, with the maximum accuracy over 10-fold cross-validation reaching 81.87%. The table below presents the results obtained by this model, including all investigated parameters; accuracy, precision, sensitivity, specificity and F1 scores (Table 8.29).

Table 8.29: The classification parameters calculated for the best performing laryngeal pathology classification systems based on laryngeal bioimpedance modality and the cutom dataset – the "small" 1D-CNN fed with Gammatone spectrograms derived from EGG speech data. "CA" stands for overall accuracy of the classifier calculated over all three classes. The first column "CI" lists classification instances in cross-validation. The following parameters were numbered according to the class, where 1 stands for cancerous and precancerous, 2 stands for neuromuscular, 3 stand for healthy. "A" is accuracy for the particular class, "P" is precision, "Sn" is sensitivity, "Sp" is specificity, and "F1" is the F1 score.

CI	CA	A1	A2	A3	P1	P2	P3	Sn1	Sn2	Sn3	Sp1	Sp2	Sp3	F1-1	F1-2	F1-3
1	79.53%	81.87%	87.91%	89.28%	78.00%	81.25%	78.73%	52.35%	85.71%	95.60%	93.96%	89.12%	85.80%	62.65%	83.42%	86.35%
2	74.36%	81.41%	79.06%	88.26%	68.75%	70.05%	83.33%	66.44%	71.98%	83.33%	87.57%	82.98%	90.94%	67.58%	71.00%	83.33%
3	68.92%	76.58%	77.70%	83.56%	64.47%	81.82%	67.44%	66.22%	42.57%	97.97%	81.76%	95.27%	76.35%	65.33%	56.00%	79.89%
4	81.87%	83.43%	87.33%	92.98%	70.51%	88.24%	85.78%	73.83%	74.18%	96.15%	87.36%	94.56%	91.24%	72.13%	80.60%	90.67%
5	79.06%	83.17%	83.17%	91.78%	68.64%	77.59%	91.07%	77.85%	74.18%	85.00%	85.36%	88.15%	95.47%	72.96%	75.84%	87.93%
6	75.70%	79.72%	90.38%	81.29%	61.95%	86.50%	82.19%	82.35%	86.07%	59.70%	78.61%	92.72%	92.99%	70.71%	86.28%	69.16%

7	69.98%	71.16%	75.65%	93.14%	59.41%	59.22%	98.28%	42.55%	86.52%	80.85%	85.46%	70.21%	99.29%	49.59%	70.32%	88.72%
8	67.13%	68.33%	91.39%	74.53%	51.76%	84.36%	63.74%	53.93%	91.28%	55.90%	75.38%	91.45%	83.94%	52.82%	87.68%	59.56%
9	67.76%	75.00%	78.97%	81.54%	61.54%	75.96%	67.96%	62.86%	54.86%	85.42%	80.90%	91.20%	79.58%	62.19%	63.71%	75.69%
10	77.82%	82.85%	79.92%	92.89%	78.81%	68.09%	87.79%	62.00%	78.05%	92.07%	92.38%	80.89%	93.31%	69.40%	72.73%	89.88%
AVERAGE	**74.21%**	**78.35%**	**83.15%**	**86.93%**	**66.38%**	**77.31%**	**80.63%**	**64.04%**	**74.54%**	**83.20%**	**84.87%**	**87.66%**	**88.89%**	**64.54%**	**74.76%**	**81.12%**
SD	**5.41**	**5.37**	**5.69**	**6.39**	**8.34**	**9.17**	**11.21**	**12.20**	**15.31**	**14.64**	**5.85**	**7.69**	**7.27**	**7.96**	**10.12**	**10.24**

During evaluation of the unimodal laryngeal condition multi-class discrimination models using SVD, the "small" 1D-CNN model fed with speech-derived GTCC matrices outperformed other methods, with the accuracy reaching a maximum of 77.78%. Furthermore, the model showed relatively strong sensitivity in detecting cancerous and precancerous lesions – on average reaching 79.98% ± 12.44, with the maximum as high as 100% (Table 8.32). The following table presents all evaluation parameters resulting from the 10-fold cross-validation completed on this model:

Table 8.30: The classification parameters calculated for the best performing laryngeal pathology classification systems based on laryngeal bioimpedance modality and SVD – the "small" 1D-CNN fed with GTCCs derived from EGG speech data. "CA" stands for overall accuracy of the classifier calculated over all three classes. The first column "CI" lists classification instances in cross-validation. The following parameters were numbered according to the class, where 1 stands for cancerous and precancerous, 2 stands for neuromuscular, 3 stand for healthy. "A" is accuracy for the particular class, "P" is precision, "Sn" is sensitivity, "Sp" is specificity, and "F1" is the F1 score.

CI	CA	A1	A2	A3	P1	P2	P3	Sn1	Sn2	Sn3	Sp1	Sp2	Sp3	F1-1	F1-2	F1-3
1	77.78%	84.72%	81.25%	89.58%	68.57%	86.21%	86.67%	100.00%	52.08%	81.25%	77.08%	95.83%	93.75%	81.36%	64.94%	83.87%
2	64.44%	70.00%	75.56%	83.33%	53.06%	78.57%	77.78%	86.67%	36.67%	70.00%	61.67%	95.00%	90.00%	65.82%	50.00%	73.68%
3	63.89%	72.22%	75.00%	80.56%	55.81%	72.73%	70.49%	80.00%	40.00%	71.67%	68.33%	92.50%	85.00%	65.75%	51.61%	71.07%
4	65.56%	73.89%	75.00%	82.22%	58.67%	67.44%	72.58%	73.33%	48.33%	75.00%	74.17%	88.33%	85.83%	65.19%	56.31%	73.77%
5	74.02%	82.84%	74.51%	90.69%	70.89%	66.00%	82.67%	82.35%	48.53%	91.18%	83.09%	87.50%	90.44%	76.19%	55.93%	86.71%
6	69.44%	78.47%	75.69%	84.72%	63.49%	74.07%	74.07%	83.33%	41.67%	83.33%	76.04%	92.71%	85.42%	72.07%	53.33%	78.43%
7	66.67%	72.78%	75.00%	85.56%	55.45%	80.00%	81.48%	93.33%	33.33%	73.33%	62.50%	95.83%	91.67%	69.57%	47.06%	77.19%
8	60.19%	69.44%	74.07%	76.85%	53.33%	72.22%	62.22%	66.67%	36.11%	77.78%	70.83%	93.06%	76.39%	59.26%	48.15%	69.14%
9	69.05%	74.40%	78.57%	85.12%	62.75%	66.13%	78.18%	57.14%	73.21%	76.79%	83.04%	81.25%	89.29%	59.81%	69.49%	77.48%
10	57.05%	62.18%	71.79%	80.13%	45.98%	78.57%	69.09%	76.92%	21.15%	73.08%	54.81%	97.12%	83.65%	57.55%	33.33%	71.03%
AVERAGE	**66.81%**	**74.10%**	**75.65%**	**83.88%**	**58.80%**	**74.19%**	**75.52%**	**79.98%**	**43.11%**	**77.34%**	**71.16%**	**91.91%**	**87.14%**	**67.26%**	**53.02%**	**76.24%**
SD	**6.14**	**6.62**	**2.58**	**4.24**	**7.65**	**6.67**	**7.28**	**12.44**	**13.84**	**6.40**	**9.38**	**4.89**	**5.00**	**7.67**	**9.95**	**5.70**

Overall, the 1D-CNNs demonstrated strong generalisation, especially for speech signal inputs, however, this architecture struggled with classification of sustained phonation, with the accuracy dropping below 55%. Both Gammatone spectrograms and GTCCs consistently outperformed Mel-spectrum derived features, confirming the conclusion of ERB-based features being more suitable for laryngeal pathology classification based on human phonation signals.

8.3.2.3. 2D-CNN Classifier

On average, the 2D-CNN model showed moderate performance (Table 8.31), with speech-derived GTCCs outperforming other feature extraction methods for the SVD dataset (64.66% ± 4.05), and raw EGG speech WAV recordings delivering the highest average accuracy in the custom dataset testing (70.15% ± 5.32). For all instances of 10-fold cross-validation, the speech signals continuously outperformed the sustained phonation, delivering the best results in a form of ERB-spectrum-based features (SVD) or raw WAV signals (custom dataset). These results suggest that while 2D-CNNs can extract meaningful spatial patterns from EGG signals, they do not surpass the temporal modelling capacity of 1D-CNNs in this application. Their strength lies in scenarios where static spatial representations dominate, but they may fall short in capturing subtle glottal dynamics across longer time windows.

Table 8.31: The accuracy of laryngeal pathology classification based on laryngeal bioimpedance (EGG), using 2D-CNN classifier.

DATASET	PHONATION TYPE	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	68.82% ± 3.31	68.71% ± 4.04	54.44% ± 13.32	69.39% ± 6.50	67.43% ± 4.81	70.15% ± 5.32
OURs	Sustained Phonation	56.34% ± 4.51	51.49% ± 7.59	45.66% ± 6.13	53.26% ± 8.53	50.06% ± 10.64	58.42% ± 6.54
SVD	Speech	59.40% ± 7.34	**64.66% ± 4.05**	60.32% ± 3.59	61.55% ± 5.79	58.67% ± 4.53	54.42% ± 5.48
SVD	Sustained Phonation	54.91% ± 3.94	53.92% ± 3.63	48.91% ± 3.79	54.19% ± 4.10	51.27% ± 3.90	53.43% ± 5.86

8.3.2.4. RNN Classifiers

The performance of RNN models in laryngeal pathology classification based on unimodal application of EGG signals can be described as low to moderate, for the most part not exceeding the range of 40-70% (Table 8.32). Both models lacked in precision and sensitivity of detecting cancerous and precancerous cases, especially when compared with the scores achieved by 1D-CNN models.

Although, on average performing significantly worse than 1D-CNN models, the trend of ERB-derived features outperforming other methodologies remained stable, with continuous speech outperforming sustained phonation for most feature extraction methods.

Table 8.32: The accuracy of laryngeal pathology classification based on laryngeal bioimpedance (EGG), using LSTM and BiLSTM classifiers.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
OURs	Speech	LSTM	60.32% ± 7.47	67.79% ± 5.36	64.09% ± 4.44	67.11% ± 5.87	58.70% ± 4.61	49.98% ± 9.54
OURs	Sustained Phonation	LSTM	50.29% ± 9.45	46.72% ± 7.30	45.27% ± 7.34	50.54% ± 9.25	48.72% ± 5.88	41.90% ± 7.79
OURs	Speech	BiLSTM	60.38% ± 5.96	69.84% ± 4.57	63.70% ± 5.13	65.29% ± 5.63	59.09% ± 4.90	49.37% ± 8.13
OURs	Sustained Phonation	BiLSTM	56.48% ± 6.10	48.54% ± 5.86	44.66% ± 6.53	55.24% ± 5.32	47.83% ± 6.46	46.62% ± 6.33
SVD	Speech	LSTM	48.48% ± 6.08	61.69% ± 5.54	52.25% ± 5.27	53.16% ± 5.99	46.01% ± 8.16	38.94% ± 5.59
SVD	Sustained Phonation	LSTM	47.83% ± 2.27	53.65% ± 3.89	48.18% ± 4.28	46.40% ± 2.90	47.14% ± 4.43	43.10% ± 5.33
SVD	Speech	BiLSTM	48.15% ± 7.65	61.53% ± 4.53	60.33% ± 3.16	57.25% ± 4.98	44.47% ± 6.65	43.83% ± 6.42
SVD	Sustained Phonation	BiLSTM	46.94% ± 3.40	53.03% ± 3.81	48.32% ± 3.29	48.29% ± 3.45	47.38% ± 4.05	42.91% ± 4.92

8.3.2.5. Conclusions on Bioimpedance-Based Unimodal Laryngeal Pathology Classification

On average, the classification of laryngeal conditions using the laryngeal bioimpedance as a sole data modality produced poorer results as compared to those achieved using audio data. The following tables (Table 8.33 and Table 8.34) show the accuracy (and its SD) of all

models designed for laryngeal pathology classification based on laryngeal bioimpedance (EGG) measurements. For both tables, the five highest accuracy parameters were highlighted in bold and in colour along with the accuracy obtained for the alternative phonation type for the particular methodology. The first table (Table 8.33) depicts the results obtained from the laryngeal pathology classification based on EGG signals obtained from the custom dataset:

Table 8.33: The accuracy of all models designed for laryngeal pathology classification with detection of cancerous and precancerous lesions based on laryngeal bioimpedance (EGG) modality performed on the custom dataset.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
EGG	Speech	RF	58.24% ± 19.56	60.21% ± 50.32	56.97% ± 24.52	60.94% ± 37.74	59.37% ± 53.42	43.96% ± 24.44
EGG	Sustained Phonation	RF	55.05% ± 45.08	51.50% ± 56.49	48.63% ± 65.37	54.27% ± 44.95	52.46% ± 44.80	44.97% ± 60.06
EGG	Speech	"small" 1D-CNN	**71.88% ± 5.03**	**72.90% ± 5.00**	71.88% ± 6.25	**74.21% ± 5.41**	67.05% ± 4.86	66.18% ± 16.91
EGG	Sustained Phonation	"small" 1D-CNN	56.19% ± 6.92	48.52% ± 7.45	43.68% ± 6.54	52.94% ± 5.49	53.06% ± 7.04	50.66% ± 12.77
EGG	Speech	"big" 1D-CNN	66.13% ± 2.96	**72.05% ± 4.43**	68.41% ± 4.01	69.70% ± 3.83	67.33% ± 3.09	**73.73% ± 5.01**
EGG	Sustained Phonation	"big" 1D-CNN	54.89% ± 12.03	48.59% ± 6.15	48.11% ± 9.91	54.12% ± 9.64	54.85% ± 8.44	57.17% ± 8.42
EGG	Speech	2D-CNN	68.82% ± 3.31	68.71% ± 4.04	54.44% ± 13.32	69.39% ± 6.50	67.43% ± 4.81	70.15% ± 5.32
EGG	Sustained Phonation	2D-CNN	56.34% ± 4.51	51.49% ± 7.59	45.66% ± 6.13	53.26% ± 8.53	50.06% ± 10.64	58.42% ± 6.54
EGG	Speech	LSTM	60.32% ± 7.47	67.79% ± 5.36	64.09% ± 4.44	67.11% ± 5.87	58.70% ± 4.61	49.98% ± 9.54
EGG	Sustained Phonation	LSTM	50.29% ± 9.45	46.72% ± 7.30	45.27% ± 7.34	50.54% ± 9.25	48.72% ± 5.88	41.90% ± 7.79
EGG	Speech	BiLSTM	60.38% ± 5.96	69.84% ± 4.57	63.70% ± 5.13	65.29% ± 5.63	59.09% ± 4.90	49.37% ± 8.13
EGG	Sustained Phonation	BiLSTM	56.48% ± 6.10	48.54% ± 5.86	44.66% ± 6.53	55.24% ± 5.32	47.83% ± 6.46	46.62% ± 6.33

The second table (Table 8.34) presents the results obtained from the laryngeal pathology classification performed on SVD:

Table 8.34: The accuracy of all models designed for laryngeal pathology classification with detection of cancerous and precancerous lesions based on laryngeal bioimpedance modality (EGG) performed on Saarbruecken Voice Database.

MODALITY	PHONATION TYPE	CLASSIFIER	STFT Spectrogram	GTCC	MFCC	Gammatone Spectrogram	Mel-Spectrogram	WAVs
EGG	Speech	RF	54.46% ± 33.09	58.07% ± 43.43	54.44% ± 20.82	59.10% ± 51.07	55.76% ± 31.05	50.65% ± 26.23
EGG	Sustained Phonation	RF	50.54% ± 25.42	53.16% ± 34.84	47.59% ± 31.34	52.26% ± 35.43	49.56% ± 28.55	42.97% ± 21.94
EGG	Speech	"small" 1D-CNN	57.42% ± 7.74	**66.81% ± 6.14**	59.13% ± 6.06	**63.21% ± 5.93**	58.35% ± 4.48	40.52% ± 9.07
EGG	Sustained Phonation	"small" 1D-CNN	50.65% ± 3.63	56.36% ± 2.21	49.46% ± 3.50	53.29% ± 3.27	52.00% ± 3.01	39.75% ± 9.17
EGG	Speech	"big" 1D-CNN	59.10% ± 4.78	**62.27% ± 4.28**	57.99% ± 7.46	61.35% ± 5.38	57.72% ± 5.26	**64.04% ± 6.08**
EGG	Sustained Phonation	"big" 1D-CNN	50.23% ± 3.80	55.28% ± 4.01	50.10% ± 3.98	53.38% ± 3.24	50.43% ± 4.01	56.19% ± 3.23
EGG	Speech	2D-CNN	59.40% ± 7.34	**64.66% ± 4.05**	60.32% ± 3.59	61.55% ± 5.79	58.67% ± 4.53	54.42% ± 5.48
EGG	Sustained Phonation	2D-CNN	54.91% ± 3.94	53.92% ± 3.63	48.91% ± 3.79	54.19% ± 4.10	51.27% ± 3.90	53.43% ± 5.86
EGG	Speech	LSTM	48.48% ± 6.08	61.69% ± 5.54	52.25% ± 5.27	53.16% ± 5.99	46.01% ± 8.16	38.94% ± 5.59
EGG	Sustained Phonation	LSTM	47.83% ± 2.27	53.65% ± 3.89	48.18% ± 4.28	46.40% ± 2.90	47.14% ± 4.43	43.10% ± 5.33
EGG	Speech	BiLSTM	48.15% ± 7.65	61.53% ± 4.53	60.33% ± 3.16	57.25% ± 4.98	44.47% ± 6.65	43.83% ± 6.42
EGG	Sustained Phonation	BiLSTM	46.94% ± 3.40	53.03% ± 3.81	48.32% ± 3.29	48.29% ± 3.45	47.38% ± 4.05	42.91% ± 4.92

In the unimodal multi-class approach to laryngeal pathology classification based solely on laryngeal bioimpedance signals (EGG), the “small” 1D-CNN model fed with continuous speech signals produced the best results for both datasets; while the custom dataset achieved the highest average accuracy in a form of Gammatone spectrograms (74.21% ± 5.41), the feature extraction method delivering the highest average accuracy for SVD was GTCC (66.81% ± 6.14). All resulting classification parameters, including accuracy, precision, sensitivity, specificity, and F1 scores calculated for the best performing EGG-based laryngeal pathology classification models can be seen in Table 8.29 (custom dataset’s speech in “small” 1D-CNN fed with Gammatone spectrograms) and Table 8.30 (SVD speech in “small” 1D-CNN fed with GTCCs).

Although the performance of speech and sustained phonation in EGG-based laryngeal pathology detection was split, showing high accuracy for both phonation types, during the multi-class discrimination of laryngeal conditions, the laryngeal bioimpedance measurements gathered during speech consistently outperformed those collected during sustained phonation. Notably, the features derived from the ERB-spectrum outperformed those of the Mel-spectrum, with raw WAV waveforms approaching similar accuracy values. To prove that continuous speech provides significantly better results in laryngeal pathology classification using laryngeal bioimpedance, the statistical significance testing was performed using ANOVA and Tukey's HSD on the best performing models for each modality:

1. For the custom dataset: speech-derived Gammatone spectrograms processed with "small" 1D CNN (accuracy of 74.21% ± 5.41) and raw WAV files of sustained phonation processed with 2D CNN (accuracy of 58.42% ± 6.54),
2. For SVD: speech-derived GTCC processed with "big" 1D CNN (accuracy of 66.81% ± 6.14) and sustained phonation-derived GTCC processed with "small" 1D CNN (accuracy of 56.36% ± 2.21).

On the custom dataset, speech signals yielded significantly higher performance than sustained phonation ($F(1,18) = 28.09$, $p = 0.00005$), with a mean difference of 0.141 and 95% CI $= [0.084, 0.196]$, accompanied by a large effect size ($\eta_p^2 = 0.610$). Likewise, on the SVD dataset, the difference remained significant ($F(1,18) = 25.65$, $p = 0.00008$), with a mean difference of 0.104 and 95% CI $= [0.061, 0.148]$, and a comparably large effect size ($\eta_p^2 = 0.588$). These results provide strong statistical evidence that continuous speech consistently conveys richer pathological information than sustained phonation for the laryngeal bioimpedance modality.

Since the major goal of the intended laryngeal pathology classification system was the accurate identification of cancerous and precancerous conditions – the most urgent class

from the medical perspective – the designed classification systems were additionally assessed based on the achieved precision and sensitivity in detecting this class of pathologies. The “small” 1D-CNN model fed with the Gammatone spectrograms derived from the speech data of the custom dataset delivered the average precision in detecting cancerous and precancerous lesions of 66.38%, while its average sensitivity reached 64.04%. With specificity (the accuracy of negative predictions) delivering an average of 84.87% for cancerous and precancerous growths, this model demonstrated the most balanced performance in identifying the malignant lesions, however, its performance as a sole modality was significantly lacking in terms of diagnostic precision.

The “small” 1D-CNN fed with SVD speech-derived GTCC matrices delivered high sensitivity of 79.98%, making the model suitable for screening applications where the priority is to avoid missed detections. Nevertheless, its precision fell below 60% (58.80%), making it unsuitable for confirmatory diagnostics of cancerous and precancerous lesions.

Overall, the results obtained from multi-class discrimination of laryngeal cases based on EGG signals suggest that while the 1D-CNN architectures exhibit promising performance, their applicability depends on the clinical context. The “small” 1D-CNN based on Gammatone spectrograms offers the best trade-off between precision and sensitivity, making it well-suited for balanced diagnostic tasks or general clinical usage. The GTCC-based “small” 1D-CNN tested on SVD demonstrates strong screening potential, prioritising high sensitivity and thus minimising missed cases. The WAV-based “big” 1D-CNN delivers more conservative cancer classification, potentially better for confirmatory diagnostics where specificity is valued over sensitivity.

These findings further confirm the utility of speech-based EGG signals in detecting malignant lesions, however, they confirm that the unimodal approach based solely on EGG signals does not deliver the precision required for accurate diagnostic classification,

especially for cancerous and precancerous lesions. They also highlight that smaller CNN architectures, when paired with informative features derived from the ERB spectrum, can outperform deeper networks for certain tasks.

Overall, the CNN-based architectures remain the most effective classifiers for the EGG signals in laryngeal pathology classification. Furthermore, the conclusion on the ERB-spectrum features, such as GTCC and Gammatone spectrograms, outperforming other feature extraction methods, including those based on Mel-spectrum, upholds for both audio and laryngeal bioimpedance signals. These results highlight the need for multimodal approaches to attain more accurate classification of malignancies.

Chapter 9

Multimodal System Results

Having investigated the laryngeal pathology classification based on audio and laryngeal bioimpedance (EGG) as a sole data input, it is clear the unimodal approach does not deliver results precise enough for medical identification of cancerous and precancerous lesions, particularly when relying on EGG as a sole data modality (chapter 8). To fulfil the major objective of this research of developing an accurate and robust laryngeal pathology classification system capable of detecting cancerous and precancerous laryngeal growths with high precision, multimodal approaches to deep learning have been investigated.

This chapter presents the results obtained from the multimodal classification experiments aimed at detecting and differentiating laryngeal pathologies using the combination of audio recordings of human phonation and simultaneous electroglottographic measurements. Building upon the unimodal findings discussed in the previous chapter, the focus here shifts to understanding how combining these complementary modalities can enhance the diagnostic potential of the developed system. The rationale for multimodal approaches lies in the hypothesis that integrating features derived from both signal types – capturing both acoustic and physiological aspects of phonatory function – provides a more comprehensive, holistic, and discriminative representation of laryngeal pathologies.

This chapter is split into two major sections according to the classification type. In the first section (section 9.1), we discuss the results obtained for the multimodal laryngeal pathology detection system – the binary classification model distinguishing between the pathological signals and control (healthy) cases. In the second section (9.2), the results of the multimodal laryngeal pathology classification system are discussed – this system

discriminates between control (healthy) cases, neuromuscular disorders, as well as the cancerous and precancerous lesions, which became the system's classification priority.

Each section is organised according to the three major fusion strategies investigated in this study – early fusion, hybrid (intermediate) fusion, and late fusion. The fusion strategies are applied following the methodology previously explored in chapter 7 of this thesis. The performance of each approach is assessed using standard evaluation metrics of accuracy, precision, sensitivity, specificity and F1 scores, all of which were introduced in section 8.1 of this thesis (*8.1. Methods of Results Assessment*). Furthermore, comparisons are made with unimodal baselines to highlight the advantages and limitations of multimodal integration.

The first part of each section (9.1.1 and 9.2.1) focuses on early fusion techniques, where audio and laryngeal bioimpedance signals, each in a form of derived features, are concatenated prior to classification. Subsections 9.1.2 and 9.2.2 explore the hybrid fusion strategy, combining modality-specific networks at an intermediate level, following into the fully connected and classification layers of the deep learning algorithm. Finally, sections 9.1.3 and 9.2.3 present the results of the late fusion approach, where the decisions from independently trained modality-specific DL models are integrated at the decision level and re-evaluated using an additional meta-classifier.

To determine which fusion strategy delivers the most reliable classification results, statistical significance testing is applied across all investigated in this chapter models – best-performing unimodal baselines (audio and laryngeal bioimpedance) and each multimodal fusion strategy – for both binary pathology detection and multi-class classification with detection of cancerous and precancerous lesions. For that, ANOVA followed by Tukey's HSD post-hoc comparisons are performed, with results reported in terms of F-statistics, p-values, confidence intervals, and effect sizes.

These experiments provide a comprehensive view of the potential and challenges of multimodal learning in the context of laryngeal pathology classification. Furthermore, the results describe the concluding best-performing multimodal laryngeal pathology detection and classification systems developed during the course of this research.

9. MULTIMODAL SYSTEM RESULTS

9.1. MULTIMODAL LARYNGEAL PATHOLOGY DETECTION

This section presents the results of binary classification experiments conducted to distinguish between healthy and pathological cases using multimodal fusion of audio signals and simultaneously recorded laryngeal bioimpedance (EGG). The design of each multimodal detection system was informed by prior unimodal evaluations, selecting the best-performing feature types and network architectures for each modality to construct effective fusion-based classifiers. For completeness, the performance of the unimodal systems was assessed across two databases: the custom dataset composed for the purposes of this study (OURs), as well as SVD – a publicly available database. The SVD dataset was included primarily for completeness and to demonstrate the generalisability of the proposed models. However, as discussed in chapter 4 of this work (*4.4.1 Limitations of SVD*), SVD has several limitations, such as repeated recordings of certain subjects and the poor representation of pathologies (low numbers of recorded samples), which reduce its reliability for robust pathology detection. As such, while results on SVD are reported, the primary performance insights and design choices are based on the custom dataset developed in this study.

For audio modality, the highest accuracy was achieved using GTCCs extracted from the continuous speech signals and processed using the “small” 1D-CNN architecture – the pathology detection model reached the average accuracy of 89.17% ± 7.30 on the custom dataset, and 80.67% ± 3.02 when tested with SVD. Alternative configurations providing robust performance in pathology detection included the “big” 1D-CNN fed with GTCCs (87.33% ± 11.07 for speech and 82.79% ± 8.58 for sustained phonation) or raw WAV files (85.42% ± 5.05 for speech and 83.77% ± 4.05 for sustained phonation), as well as BiLSTMs

fed with GTCCs (84.71% ± 11.36 for speech and 79.31% ± 4.59 for sustained phonation). For all best-performing classification instances, speech outperformed sustained phonation.

For laryngeal bioimpedance signals in laryngeal pathology detection, the classification accuracy ratio of continuous speech against that achieved by sustained phonation was more balanced, which complicated the selection of the best-performing phonation type for the multimodal application. While sustained phonation-derived STFT spectrograms fed into a 2D-CNN model achieved the highest accuracy in the custom dataset (87.39% ± 2.50), the speech-based signals demonstrated superior generalisation consistently outperforming sustained phonation on the SVD dataset. The highest average accuracy achieved by EGG signals from SVD was 80.99% ± 1.35 (speech-derived GTCCs fed into the “small” 1D-CNN – when tested with sustained phonation, the model reached 72.59% ± 3.45). Furthermore, continuous speech proved more effective in multi-class classification (as detailed in section 9.2). Accordingly, speech was selected as the input modality for laryngeal bioimpedance in all multimodal configurations. The top-performing EGG configurations on the custom dataset included Gammatone spectrograms fed into the “small” and the “big” 1D-CNNs (84.83% ± 6.23 and 84.56% ± 4.50), as well as raw WAVs evaluated with a 2D-CNN (86.03% ± 4.94).

Summarising, the phonation type of continuous speech was chosen for both data modalities considered in this research – audio and laryngeal bioimpedance. The features were derived accordingly and processed following the methodology outlined in chapter 7 of this work. The following sections describe the results obtained for the multimodal laryngeal pathology detection systems, based on the combined input of both modalities. The three fusion strategies are considered – early, hybrid (intermediate), and late – each constructed using the optimal modality-specific configurations. The models’ performance is evaluated in terms of accuracy, precision, sensitivity, specificity, and F1 score, with comparisons drawn against the best-performing unimodal baselines.

9.1.1 Early Fusion for Laryngeal Pathology Detection

The details of the early fusion methodology implemented in this research are available in section 7.1 of this thesis (*7.1. Early Fusion*).

In the early fusion approach, the features from both modalities were concatenated prior to the classification processes. For audio signals, the GTCCs were chosen as the feature extraction method, while the laryngeal bioimpedance was processed in the form of Gammatone spectrograms. The concatenated feature matrices were subsequently fed into the "small" 1D-CNN classification model. The following figure shows the average confusion matrices achieved for the laryngeal pathology detection using the early fusion strategy, calculated over 10-fold cross-validation performed on the custom dataset (Figure 9.1A), as well as SVD (Figure 9.1B).

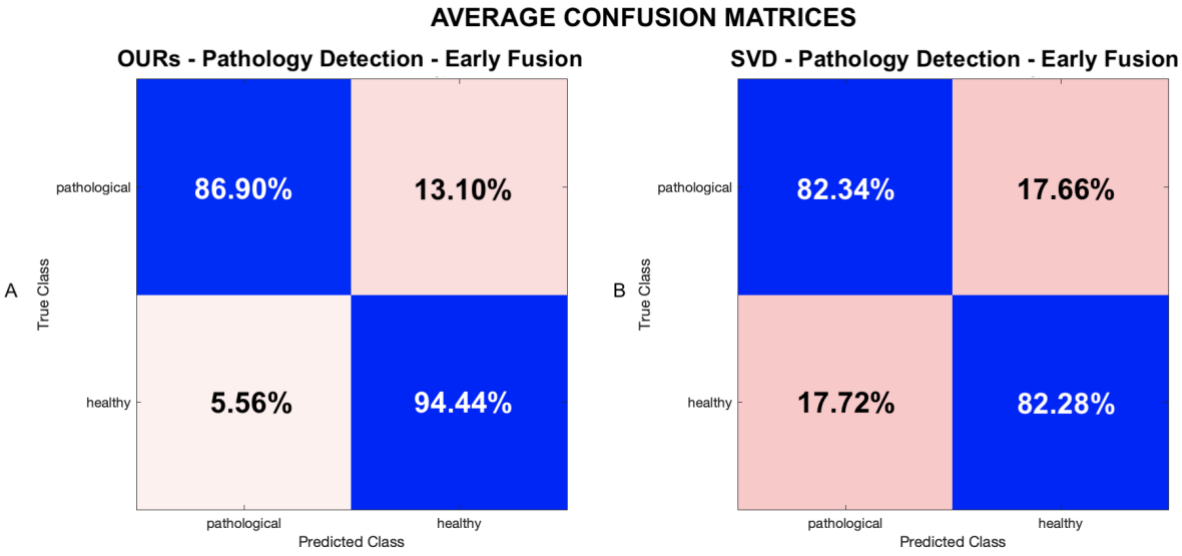

Figure 9.1: The average confusion matrices obtained for the designed early fusion multimodal system for the laryngeal pathology detection, tested over 10-fold cross-validation on the custom dataset (figure A) and SVD (figure B).

These results confirmed the effectiveness of this approach, with a notable boost in classification accuracy compared to individual modalities alone. When tested on the custom dataset, the early fusion model achieved an overall accuracy of 90.67% ± 5.34, outperforming both unimodal baselines. Precision, specificity, and F1 scores were notably

high (93.87% ± 3.09, 94.44% ± 2.59, and 90.12% ± 5.95, respectively), indicating the model's strong ability to correctly identify pathological and healthy cases alike. Nevertheless, the early fusion model's sensitivity (86.90% ± 8.81) fell slightly below the audio-based unimodal system for the laryngeal pathology detection (87.88% ± 8.66).

The following two figures represent the average accuracy, precision, sensitivity, specificity, and F1 scores achieved using the early fusion strategy on both datasets. Figure 9.2 shows the results of the custom dataset and 9.3 shows the parameters calculated for SVD testing.

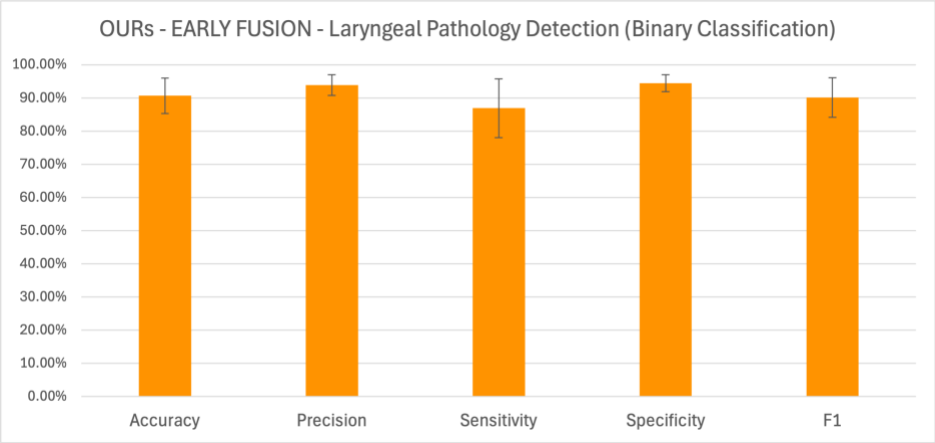

Figure 9.2: Early Fusion Model in Laryngeal Pathology Detection – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing.

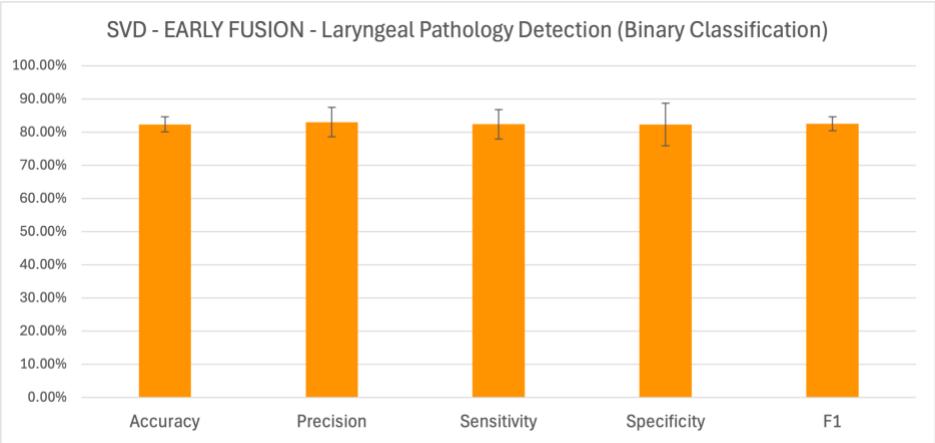

Figure 9.3: Early Fusion Model in Laryngeal Pathology Detection – the accuracy, precision, sensitivity, specificity and F1 scores calculated for SVD data testing.

9.1.2 Hybrid (Intermediate) Fusion for Laryngeal Pathology Detection

The details of the hybrid (intermediate) fusion methodology implemented in this research are available in section 7.2 of this thesis (*7.2. Hybrid (Intermediate) Fusion*).

The hybrid fusion multimodal system was designed to allow each modality to be processed independently before merging their learned representations to subsequently follow into the fully connected and classification layers. In this model, the audio signals were processed as GTCCs and fed into the 1D-CNN architecture. The laryngeal bioimpedance signals were passed through a 2D-CNN architecture, tested in two alternative forms: the raw waveform (WAV format), as well as Gammatone spectrograms. This decision stemmed from the observation that these feature types (GTCC for audio, Gammatone spectrograms and WAVs especially for EGG) processed through the chosen architectures (1D-CNN for audio and 2D-CNN for EGG) demonstrated the highest standalone performance in the unimodal evaluations.

Each modality-specific subnetwork concluded with global average pooling, followed by the fully connected layer and the flatten layer. Subsequently, the intermediate features (the representations processed by the corresponding modality-specific branches) were concatenated at the concatenation layer, following into the next two fully connected layers interleaved with a ReLU activation and a dropout layer of 20% dropout rate. Finally, the fused network followed into the softmax function and the final classification layer for the prediction.

The following figure shows the average confusion matrices obtained for the laryngeal pathology detection using the hybrid fusion multimodal approach, calculated over the 10-fold cross-validation (Figure 9.4). The confusion matrices on the left side of the figure (Figure 9.4A and Figure 9.4C) were obtained for EGG fed into the hybrid model as WAV files, and

the confusion matrices on the right (Figure 9.4B and Figure 9.4D) were obtained for EGG as Gammatone spectrograms.

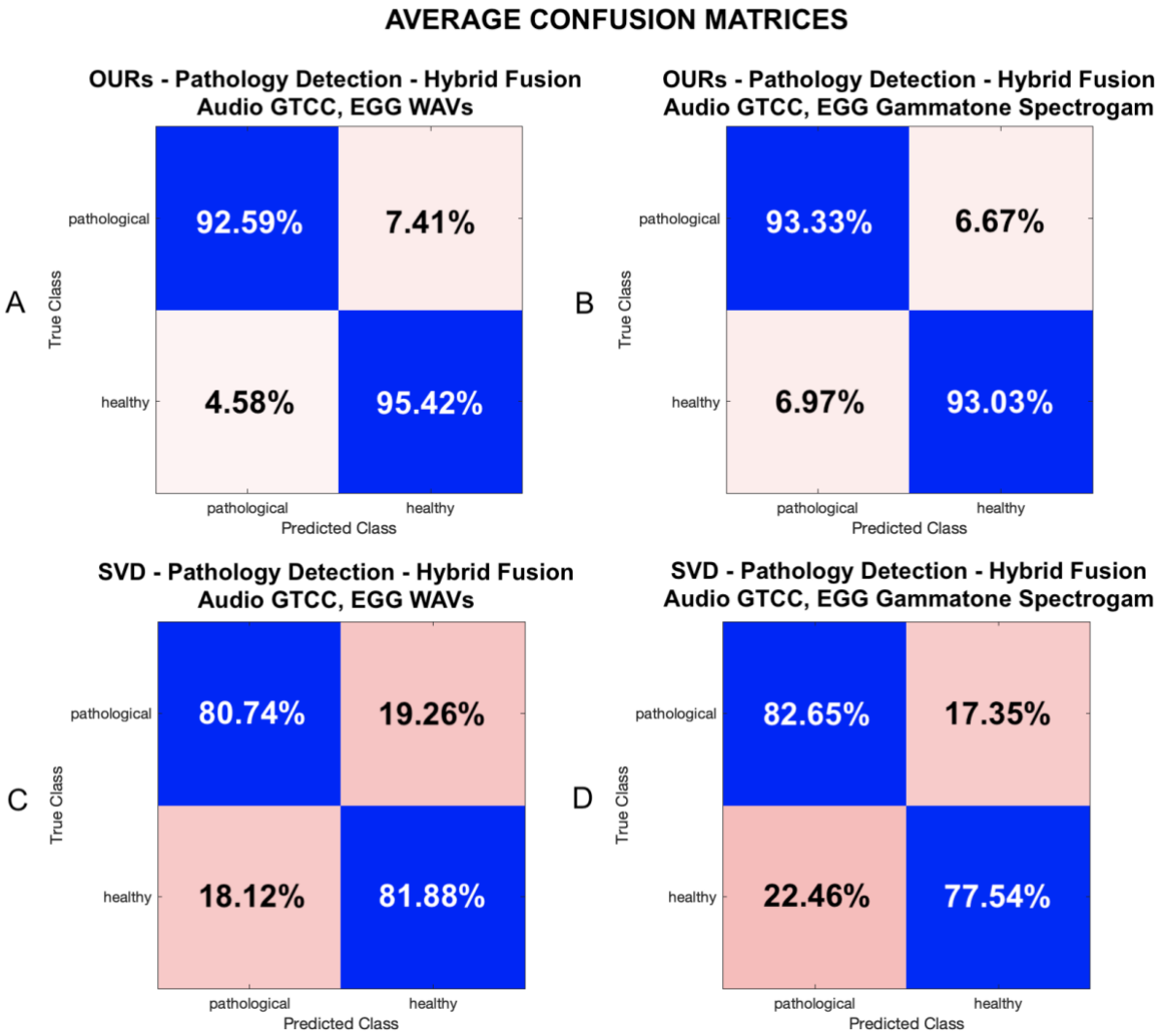

Figure 9.4: The average confusion matrices obtained for the designed hybrid fusion multimodal system for the laryngeal pathology detection, tested over 10-fold cross-validation on the custom dataset (figure A and B) and SVD (figure C and D). A and C present the confusion matrices calculated for EGG signals fed into the model as WAV files, and B and D show the confusion matrices calculated for EGG signals fed into the model as Gammatone spectrograms.

Based on the accuracy, precision, and F1 score calculated over 10-fold cross-validation of the designed hybrid models on the custom dataset, the combination of audio-derived GTCCs and raw WAV recordings of the laryngeal bioimpedance delivered the best performance (94.00% ± 4.46% of accuracy, 95.51% ± 6.11 of precision, and 93.98% ± 4.34% of F1 score). The model fed with Gammatone spectrograms derived from the

bioimpedance signals delivered a higher sensitivity: 93.07% ± 3.53%. Both models significantly outperformed the unimodal approaches to laryngeal pathology classification. Furthermore, the hybrid fusion strategy outperformed that of the early fusion. However, the results obtained during the hybrid fusion models' testing on SVD provided similar results to those achieved by the unimodal systems (an accuracy of 81.30% ± 2.68 for audio and 80.99% ± 1.35% for bioimpedance), and early fusion.

The below figures show the average accuracy, precision, sensitivity, specificity, and F1 scores achieved using the two hybrid fusion models on both datasets, where: Figure 9.5 shows the results for the custom dataset with EGG as WAVs; Figure 9.6 shows the results for the custom data with EGG as Gammatone spectrograms; Figure 9.7 presents the results for SVD with EGG as WAVs and 9.8 depicts the parameters calculated for SVD with EGG as Gammatone spectrograms.

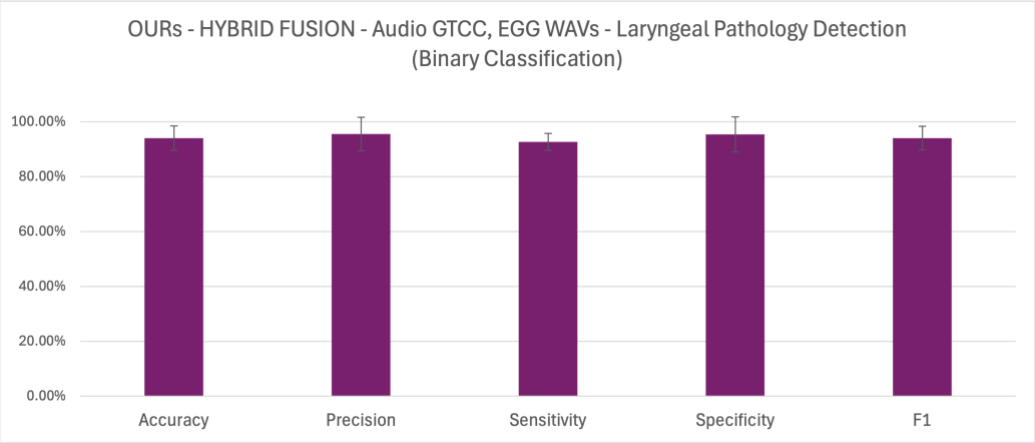

Figure 9.5: Hybrid Fusion Model in Laryngeal Pathology Detection fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing.

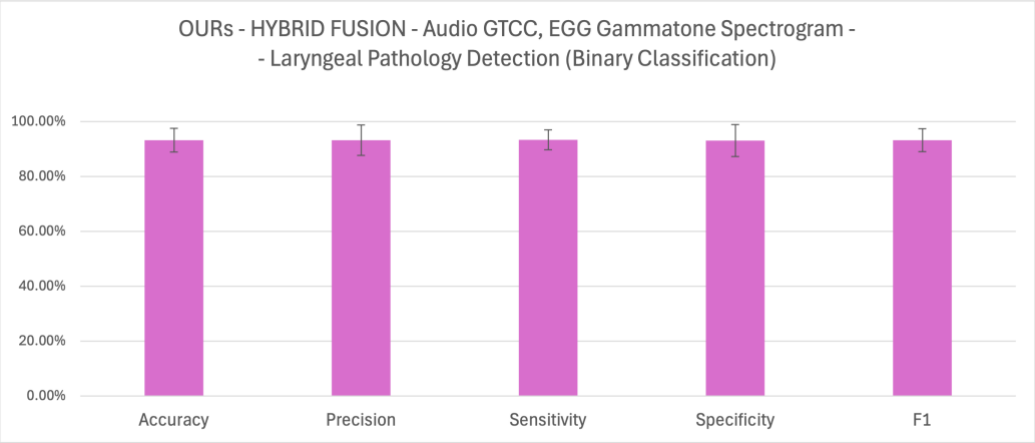

Figure 9.6: Hybrid Fusion Model in Laryngeal Pathology Detection fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing.

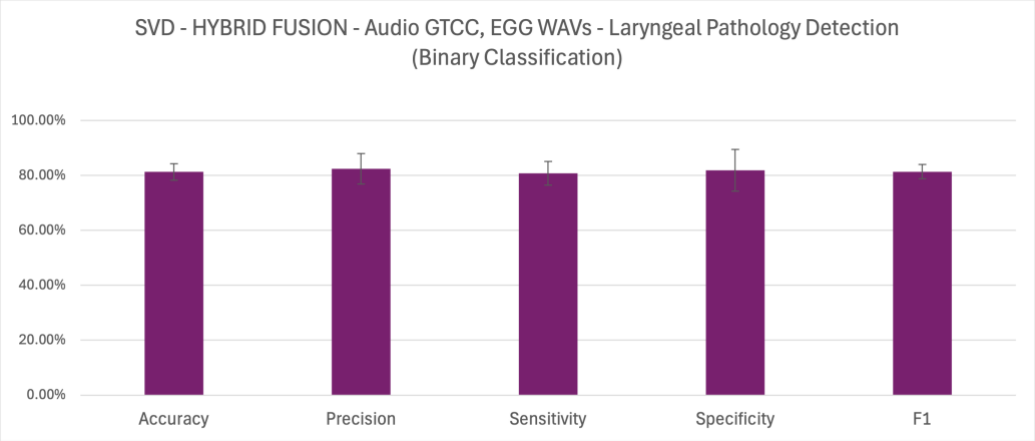

Figure 9.7: Hybrid Fusion Model in Laryngeal Pathology Detection fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing.

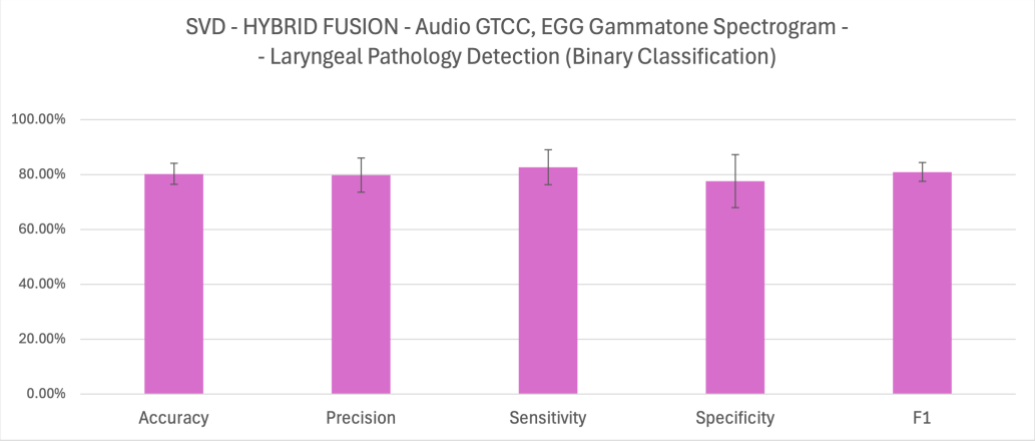

Figure 9.8: Hybrid Fusion Model in Laryngeal Pathology Detection fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing.

For the custom dataset, the hybrid strategy achieved robust results, outperforming the designed unimodal systems, and offering a flexible fusion mechanism with minimal preprocessing for the laryngeal bioimpedance modality. The achieved results mark a substantial performance gain over both unimodal and early fusion systems, suggesting that preserving modality-specific preprocessing and network structures allows for deeper and more informative representations prior to fusion.

9.1.3 Late Fusion for Laryngeal Pathology Detection

The details of the late fusion methodology implemented in this research are available in section 7.3 of this thesis (*7.3. Late Fusion*).

The late fusion experiments aimed to compare the effectiveness of combining modality-specific classifiers at the decision level. The outputs of the two independent modality-specific DL classifiers were integrated using a stacked generalisation technique. The ensemble-based approach allowed each modality to contribute a separate prediction, which was then combined by a higher-level meta-classifier.

Two distinct late fusion configurations were examined:

1. Audio speech-derived GTCCs processed using the “small” 1D-CNN, combined with bioimpedance speech-derived Gammatone spectrograms processed using the “small” 1D-CNN.
2. Audio speech-derived GTCCs processed using the “small” 1D-CNN, combined with raw bioimpedance speech WAVs processed using the “big” 1D-CNN.

The above were selected based on the results obtained from the best-performing unimodal laryngeal pathology detection systems, discussed previously in chapter 8 of this thesis. Testing of the two configurations enabled the comparison of the bioimpedance feature representation and its impact on the multimodal classification accuracy. The first

configuration preserved a parallel network structure – both modalities using spectrogram-based features on lightweight networks. The second configuration explored a more diverse setup, where the simple WAV format representation of laryngeal bioimpedance signals was paired with a deeper network capable of modelling complex temporal structures.

In both cases, the modality-specific classifiers were trained independently to learn robust unimodal feature representations, and their predicted class probabilities were subsequently used as the input for the meta-classifier. The designed meta-classifier follows the stacking ECOC-based method, and it was trained to deliver the optimal decision boundaries based on the combined output probabilities. Although primarily designed for multi-class problems, in this study the ECOC-based meta-classifier was applied in both detection (binary classification) as well as the multi-class discrimination of the pathologies.

The following figure represents the confusion matrices calculated for the late fusion multimodal laryngeal pathology detection system over the 10-fold cross-validation approach (Figure 9.9). The confusion matrices on the left side of the figure (Figure 9.9A and Figure 9.9C) were obtained for EGG fed into the hybrid model as WAV files, and the confusion matrices on the right (Figure 9.9B and Figure 9.9D) were obtained using EGG-derived Gammatone spectrograms.

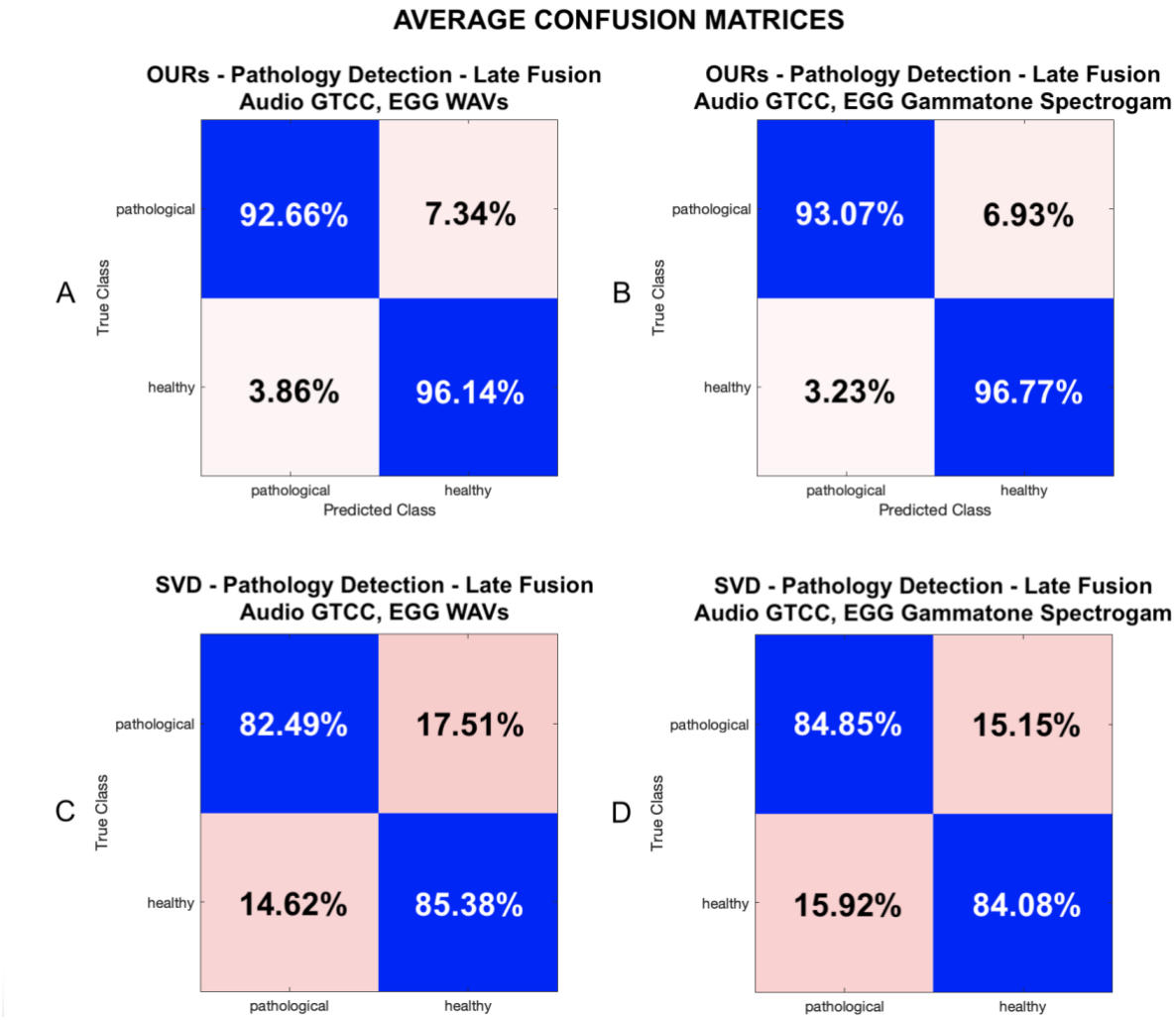

Figure 9.9: The average confusion matrices obtained for the designed late fusion multimodal system for laryngeal pathology detection, tested over 10-fold cross-validation on the custom dataset (figure A and B) and SVD (figure C and D). A and C present the confusion matrices calculated for EGG signals fed into the model as WAV files, and B and D show the confusion matrices calculated for EGG signals fed into the model as Gammatone spectrograms.

Based on the results obtained from the custom dataset testing, the first approach – using EGG-derived Gammatone spectrograms processed with the "small" 1D-CNN – demonstrated superior performance, achieving the highest accuracy, precision, specificity and F1 scores (94.92% ± 2.82, 96.67% ± 2.90%, 96.77% ± 2.84, and 94.81% ± 2.90, respectively) among all tested classification systems. The designed late fusion system outperformed both the unimodal systems, as well as other multimodal fusion strategies.

Furthermore, the results obtained during the SVD testing confirmed the conclusions drawn from the custom dataset testing – the accuracy, sensitivity, and F1 scores were the

highest for the late fusion approach using Gammatone spectrograms as the chosen feature extraction method for the laryngeal bioimpedance signals (84.46% ± 2.61, 84.85% ± 3.03, and 84.66% ± 2.63, respectively). However, the second late fusion multimodal system, while tested on SVD, achieved the highest precision and specificity among all SVD-tested laryngeal pathology detection systems (precision of 85.25% ± 2.30, and specificity of 85.38% ± 2.24).

The following figures depict the average accuracy, precision, sensitivity, specificity, and F1 scores achieved using the two late fusion multimodal pathology detection systems tested on both datasets. Furthermore, the figures include the average accuracy of the unimodal audio-based and bioimpedance-based pathology detection systems used during the late fusion, for completeness. Figure 9.10 shows the results for the custom dataset with EGG as WAVs, Figure 9.11 shows the results for the custom data with EGG as Gammatone spectrograms, Figure 9.12 presents the results for SVD with EGG as WAVs, and 9.13 depicts the results calculated for SVD with EGG as Gammatone spectrograms.

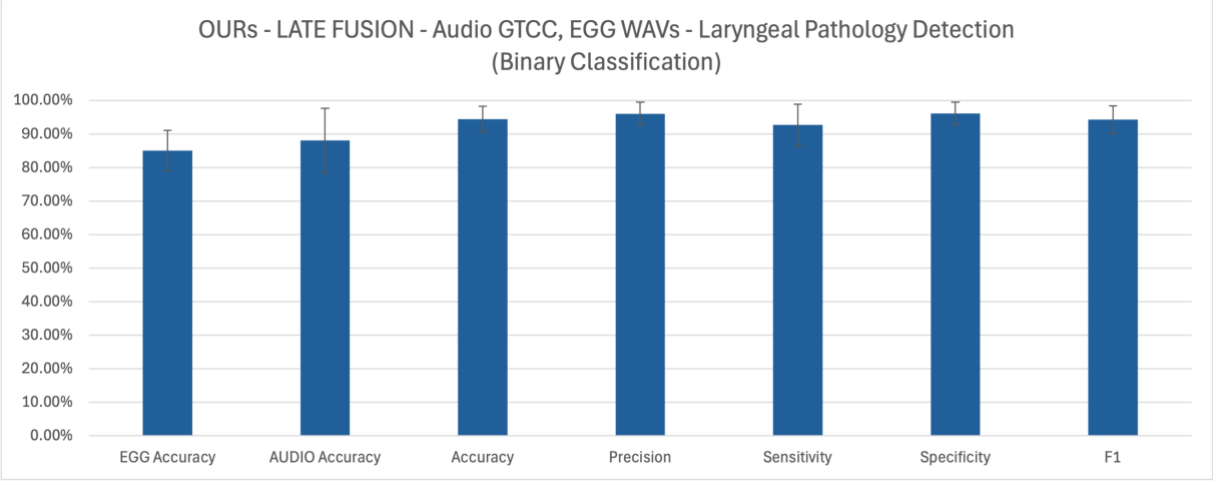

Figure 9.10: Late Fusion Model in Laryngeal Pathology Detection fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing.

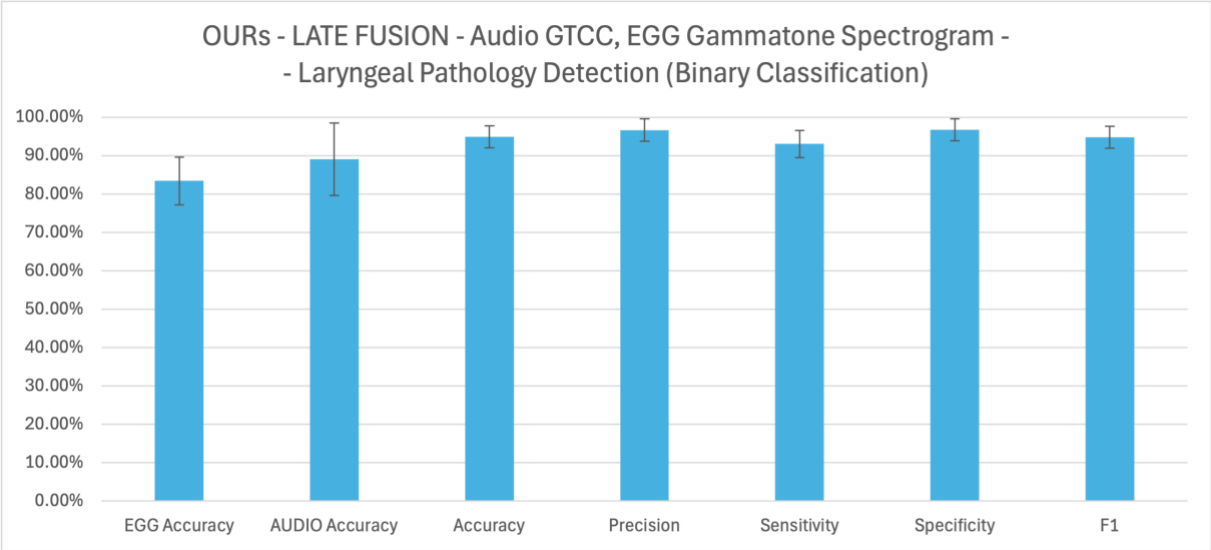

Figure 9.11: Late Fusion Model in Laryngeal Pathology Detection fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing.

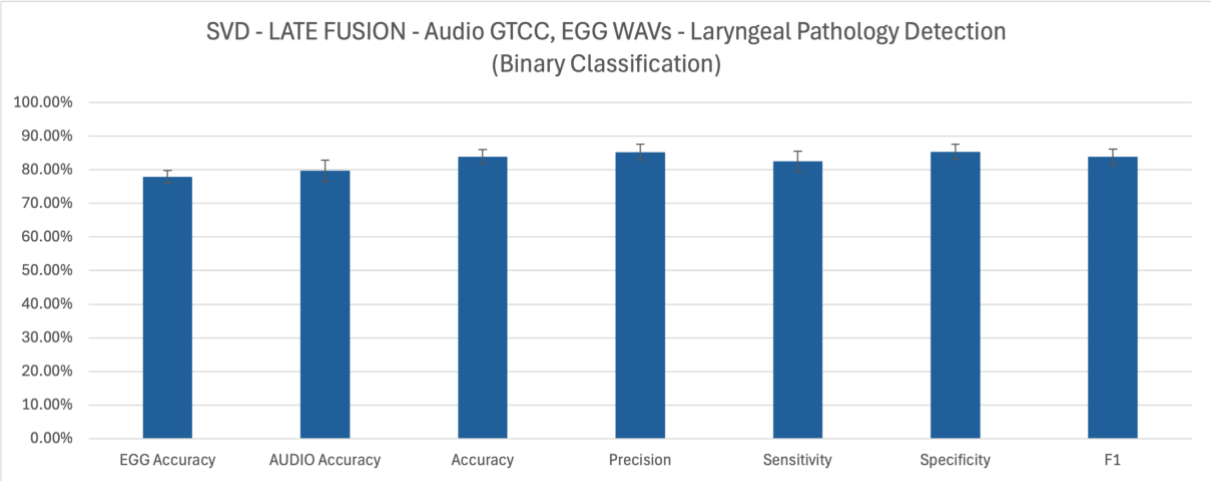

Figure 9.12: Late Fusion Model in Laryngeal Pathology Detection fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing.

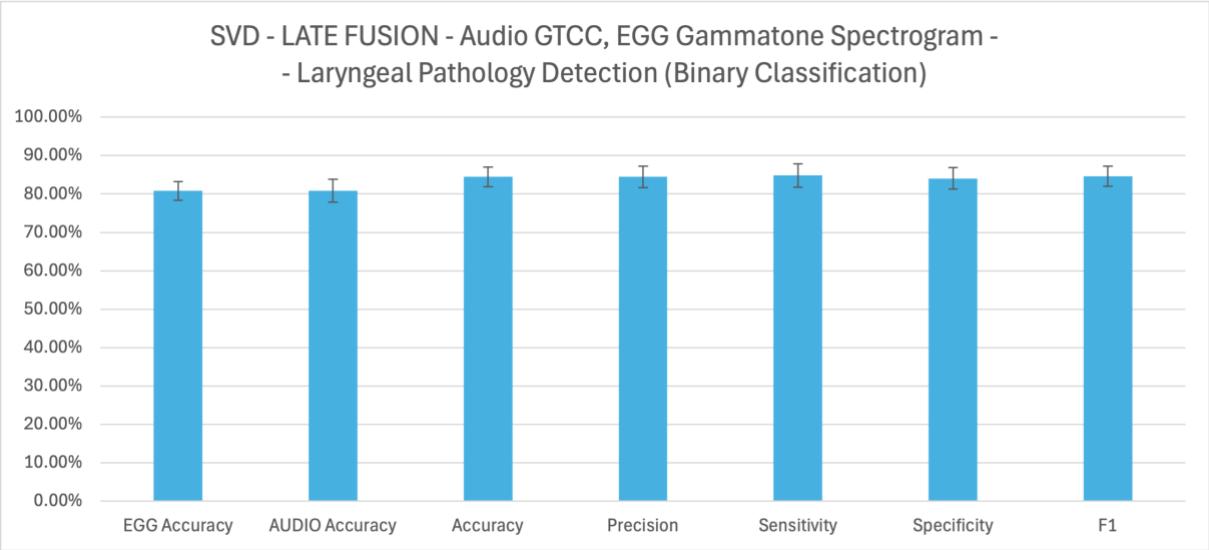

Figure 9.13: Hybrid Fusion Model in Laryngeal Pathology Detection fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing

These results reinforce the observation that the appropriate fusion of independently processed audio and laryngeal bioimpedance modalities paired with robust DL architectures can retain diagnostic information that might otherwise be lost in feature concatenation. Furthermore, the late fusion framework offers flexibility in model design, allowing each branch to be optimised independently before integration. The use of a meta-classifier allowed the system to effectively learn how to weigh the predictions from each modality, making late fusion the most accurate and balanced fusion strategy for binary pathology detection.

Overall, stacked generalisation proved to be an effective fusion strategy, particularly in cases where modalities differ significantly in their preprocessing requirements. It capitalised on the individual strengths of each modality-specific network while enabling a robust and accurate integration at the decision level.

9.1.4 Conclusions on Multimodal Laryngeal Pathology Detection

To conclude the laryngeal pathology detection analysis (binary classification between the control and pathological signals), a comparative evaluation of the best performing unimodal

and all multimodal systems was conducted across five performance metrics: accuracy, precision, sensitivity, specificity, and F1-score. Additionally, the statistical significance testing was completed on the obtained results to confirm the best-performing model. The following tables represent the average performance metrics calculated over the 10-fold cross-validation performed on the best performing unimodal systems and all multimodal classifiers designed for the purposes of laryngeal pathology detection – Table 9.1 shows the results of the custom dataset testing, while Table 9.2 depicts the values of performance metrics achieved by SVD.

Table 9.1: Performance metrics calculated for the best performing unimodal systems and all designed multimodal systems designed for the purposes of the laryngeal pathology detection (binary classification) tested using the custom dataset.

OURs – DETECTION – METHOD	Accuracy	Precision	Sensitivity	Specificity	F1
BEST UNIMODAL on AUDIO - GTCC on "small" 1D-CNN	89.17% ± 7.30	90.52% ± 7.94	87.88% ± 8.66	90.46% ± 8.61	88.99% ± 7.39
BEST UNIMODAL on EGG - WAVs on 2D-CNN	86.03% ± 4.94	92.98% ± 5.20	77.99% ± 7.35	94.06% ± 4.42	84.69% ± 5.65
EARLY FUSION - speech small 1D CNN - Pathology vs Healthy	90.67% ± 5.34	93.87% ± 3.09	86.90% ± 8.81	94.44% ± 2.59	90.12% ± 5.95
HYBRID - audio GTCC small CNN1, EGG wav CNN2	94.00% ± 4.46	95.51% ± 6.11	92.59% ± 3.10	95.42% ± 6.36	93.98% ± 4.34
HYBRID - audio GTCC small CNN1, EGG Gammatone Spectr CNN2	93.18% ± 4.27	93.23% ± 5.56	93.33% ± 3.58	93.03% ± 5.83	93.23% ± 4.16
LATE FUSION STACKED - audio GTCC, EGG WAV - small 1D CNN	94.40% ± 3.87	96.04% ± 3.38	92.66% ± 6.17	96.14% ± 3.31	94.23% ± 4.09
LATE FUSION STACKED - audio GTCC, EGG GammaSpec - small 1D CNN	94.92% ± 2.82	96.67% ± 2.90	93.07% ± 3.53	96.77% ± 2.84	94.81% ± 2.90

Table 9.2: Performance metrics calculated for the best performing unimodal systems and all designed multimodal systems designed for the purposes of the laryngeal pathology detection (binary classification) tested using SVD.

SVD – DETECTION – METHOD	Accuracy	Precision	Sensitivity	Specificity	F1
BEST UNIMODAL on AUDIO - Gammatone Spectrograms on "small" 1D-CNN	81.30% ± 2.68	82.73% ± 3.10	79.76% ± 4.68	82.88% ± 3.52	81.14% ± 3.08
BEST UNIMODAL on EGG - GTCC on "small" 1D-CNN	80.99% ± 1.35	80.19% ± 3.04	83.22% ± 3.98	78.63% ± 4.97	81.56% ± 1.45
EARLY FUSION - speech small 1D CNN - Pathology vs Healthy	82.33% ± 2.25	82.99% ± 4.41	82.34% ± 4.42	82.28% ± 6.43	82.49% ± 2.09
HYBRID - audio GTCC small CNN1, EGG wav CNN2	81.26% ± 3.02	82.41% ± 5.51	80.74% ± 4.32	81.88% ± 7.55	81.36% ± 2.62

HYBRID - audio GTCC small CNN1, EGG Gammatone Spectr CNN2	80.21% ± 3.81	79.76% ± 6.20	82.65% ± 6.37	77.54% ± 9.61	80.86% ± 3.42
LATE FUSION STACKED - audio GTCC, EGG WAV - small 1D CNN	83.92% ± 2.09	85.25% ± 2.30	82.49% ± 3.05	85.38% ± 2.24	83.82% ± 2.26
LATE FUSION STACKED - audio GTCC, EGG GammaSpec - small 1D CNN	84.46% ± 2.61	84.51% ± 2.80	84.85% ± 3.03	84.08% ± 2.85	84.66% ± 2.63

These results are also summarised in the following two figures, which visually compare all parameters calculated for the models – Figure 9.14 shows the results obtained for the custom dataset testing, while Figure 9.15 depicts the results calculated for the SVD testing. The group numbers have been assigned to the corresponding models as follows:

- 1: BEST UNIMODAL on AUDIO (for custom dataset: GTCC on “small” 1D-CNN; for SVD: Gammatone Spectrograms on “small” 1D-CNN),
- 2: BEST UNIMODAL on EGG (for custom dataset: WAVs on 2D-CNN; for SVD: GTCC on “small” 1D-CNN),
- 3: EARLY FUSION (audio GTCC and EGG Gammatone spectrograms on “small” 1D-CNN),
- 4: HYBRID (audio GTCC on “small” 1D-CNN, EGG WAVs on 2D-CNN),
- 5: HYBRID (audio GTCC on “small” 1D-CNN, EGG Gammatone Spectrograms on 2D-CNN),
- 6: LATE FUSION STACKED (audio GTCC and EGG WAVs on “small” 1D-CNN),
- 7: LATE FUSION STACKED (audio GTCC and EGG Gammatone Spectrograms on “small” 1D-CNN).

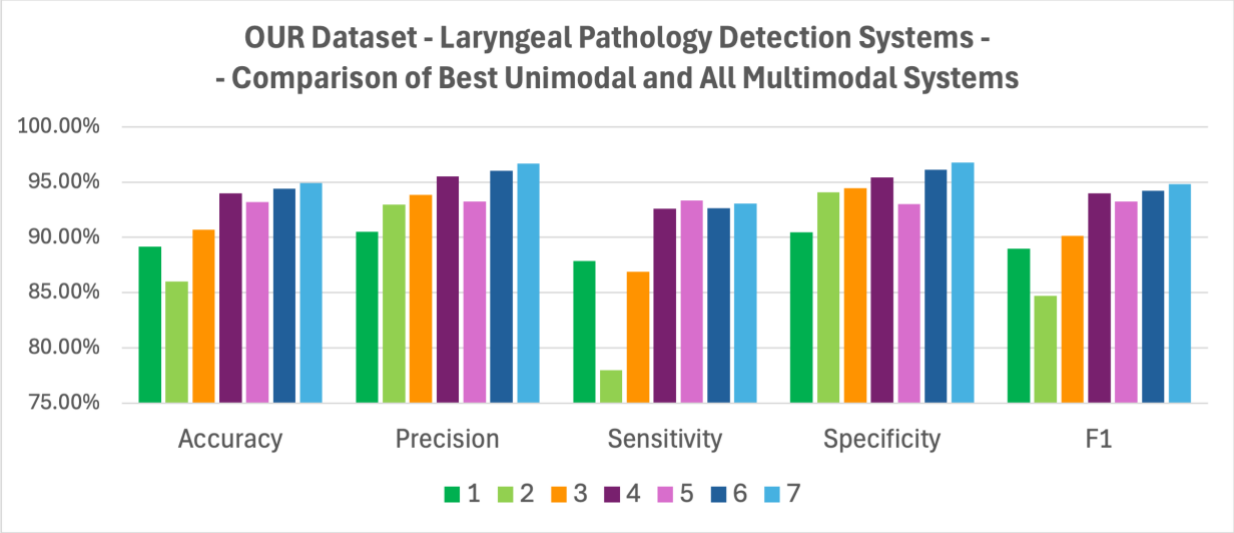

Figure 9.14: Visual representation of the results obtained for the best performing unimodal laryngeal pathology detection systems and all multimodal systems designed, tested on the custom dataset, depicting the accuracy, precision, sensitivity, specificity, and F1 score parameters.

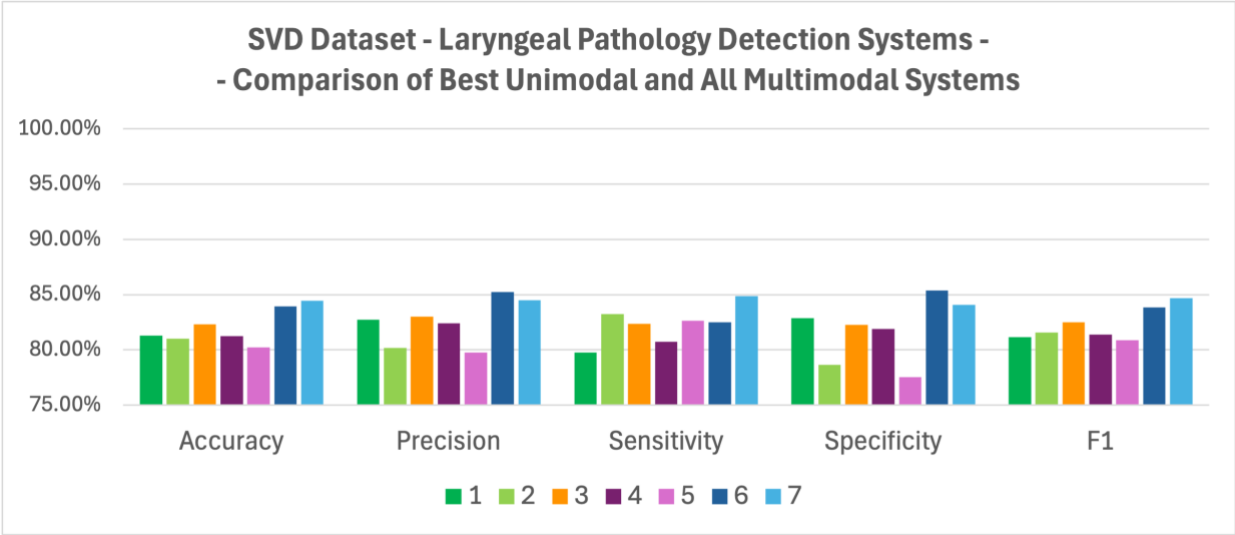

Figure 9.15: Visual representation of the results obtained for the best performing unimodal laryngeal pathology detection systems and all multimodal systems designed, tested on SVD, depicting the accuracy, precision, sensitivity, specificity, and F1 score parameters.

As shown in the chart, the best-performing system overall was the late fusion stacked model combining GTCC features from audio with Gammatone spectrograms derived from the laryngeal bioimpedance (model 7). This configuration achieved the highest accuracy on both datasets (94.92% for the custom dataset and 84.46% for SVD), as well as the highest precision for the custom dataset (96.67%) and sensitivity for SVD (84.85%) – metrics of

particular importance in medical screening contexts, where false negatives must be minimised. The highest F1 scores of 94.81% (OURs dataset) and 84.66% (SVD) achieved by this late fusion multimodal system further highlight the model's balanced performance across positive and negative classes.

The hybrid fusion system based on audio-derived GTCCs and raw EGG WAVs processed through independent 1D-CNN branches (model 4) also delivered excellent performance with an accuracy of 94.00%, precision of 95.51%, and F1-score of 93.98%. While slightly below the late fusion model, its comparable results validate the strength of intermediate-level feature integration. Nevertheless, this model underperformed on the SVD testing, delivering the results similar to those obtained using the unimodal and early fusion models.

The early fusion approach (model 3), though simpler in design, still outperformed both unimodal baselines, achieving an accuracy of 90.67% on the custom dataset. This demonstrates that even straightforward fusion at the feature level can effectively leverage the complementary information from both modalities.

Unimodal models, while effective in isolation, showed limitations in either sensitivity or specificity. The audio GTCC model (model 1) tested on the custom dataset provided a good balance (accuracy: 89.17%, sensitivity: 87.88%), while the EGG WAV model (model 2) offered higher precision (92.98%) but notably lower sensitivity (77.99%), highlighting the value of combining these modalities to compensate for individual weaknesses.

To formally assess whether the observed accuracy differences between unimodal and multimodal systems were statistically significant, one-way ANOVA followed by Tukey's HSD post-hoc comparisons was conducted for both datasets.

For the custom dataset, the ANOVA testing indicated a significant main effect of model type on detection accuracy, $F(6,63) = 4.51$, $p < 0.001$ ($p = 0.0007$), with the effect size

(partial eta squared) of $\eta_p^2 = 0.301$, suggesting that approximately 30% of the variance in performance could be attributed to the choice of classifier. Post-hoc analysis revealed that the late fusion stacked models (models 6 and 7) significantly outperformed the unimodal EGG baseline (model 2), with mean accuracy differences ranging from -0.084 to -0.089 (95% CI) and $p < 0.005$. Similarly, both of the hybrid fusion systems (models 4 and 5) achieved significantly higher accuracy than the unimodal EGG system (mean difference -0.072, 95% CI $= [-0.138, -0.005]$, $p < 0.028$ for model 2 and 5, and mean difference -0.080, 95% CI $= [-0.146, -0.013]$, $p < 0.009$ for model 2 and 4). Nevertheless, no significant differences were observed between the audio-only baseline (model 1) and any of the multimodal models (all $p > 0.13$), indicating that the largest gains came from overcoming the limitations of the EGG-only model.

For SVD, ANOVA again showed a significant effect of model type on detection accuracy, $F(6,63) = 3.61$, $p < 0.004$, with the effect size (partial eta squared) of $\eta_p^2 = 0.256$, reflecting a medium-to-large effect. Tukey's HSD revealed that the late fusion stacked model based on audio-derived GTCC with EGG-derived Gammatone spectrograms (model 7) significantly outperformed the hybrid fusion model using audio-derived GTCCs combined with EGG-derived Gammatone spectrograms (model 5), with a mean difference of -0.043 (95% CI $= [-0.079, -0.007]$, $p < 0.011$). However, at this level of classification, no significant differences were observed between the unimodal baselines and the best multimodal systems (all $p > 0.07$), suggesting that the improvements of multimodal fusion were more modest on SVD compared to the custom dataset.

Taken together, these results confirm that multimodal fusion – particularly late fusion approaches – produced reliable gains in pathology detection performance. While audio alone remained a strong modality, the combination with EGG features consistently reduced error rates, especially by offsetting the weaknesses of the unimodal EGG classifier. The

effect sizes indicate that the influence of fusion strategy on model accuracy was substantial, further supporting the conclusion that multimodal integration enhances system robustness in binary pathology detection.

In summary, the multimodal fusion systems – particularly the late model – not only improved overall detection performance but also enhanced the system's reliability in identifying pathological signals. These findings strongly support the integration of both audio and laryngeal bioimpedance signals in clinical diagnostic tools, offering a more robust and generalisable solution for automatic voice pathology detection.

9.2. MULTIMODAL LARYNGEAL PATHOLOGY CLASSIFICATION

In this section, we present and discuss the results obtained from the multimodal multi-class laryngeal pathology classification systems developed for the purposes of this study using early, hybrid (intermediate), and late fusion approaches. The aim was to construct the most accurate fusion-based model leveraging the complementary information from audio signals and simultaneously recorded laryngeal bioimpedance for multi-class discrimination between cancerous and precancerous growths, neuromuscular disorders, and control (healthy) cases. The priority during the development of the intended system was its ability to accurately identify malignancies (the cancerous and precancerous lesions) providing the highest precision and sensitivity.

The methodology adopted for the development of the envisaged multimodal laryngeal pathology classification system closely followed the approach used in the pathology detection (section 9.1), whereby the classification architecture and the feature choices were directly informed by the performance of the unimodal systems developed for the same

purpose. Since the multi-class setting presents a more complex classification landscape, the selection of robust and generalisable feature representations was particularly critical.

For the generalisability purposes, all developed multimodal laryngeal pathology classification models were evaluated on two independent datasets: the custom dataset developed for the purposes of this study, as well as SVD. It is important to note that SVD was included primarily for completeness and validation purposes, since the dataset itself exhibits several limitations such as poor pathology representation, particularly in case of the malignant cases. As discussed in section 4.4 of this work (*4.4.1 Limitations of SVD*), SVD's limitations affect its reliability as a benchmark dataset for the task of laryngeal pathology classification capable of accurately identifying cancerous lesions. Therefore, while the results obtained from the SVD testing are reported and discussed, the core design decisions and performance interpretations are grounded in the custom dataset developed during the course of this research.

Unlike the case of laryngeal bioimpedance signals in the pathology detection system, the key observation underpinning the multi-class discrimination is that continuous speech recordings significantly outperformed the sustained phonation across both audio and laryngeal bioimpedance modalities. This trend was consistently observed across both the custom dataset as well as SVD. As a result, all multimodal systems for multi-class classification were exclusively developed using speech signals. Sustained phonation was excluded at this stage to avoid diluting classification accuracy with suboptimal input data.

In unimodal classification testing completed on audio modality, the best performance was achieved using GTCCs processed using the "small" 1D-CNN model, reaching the average accuracy over 10-fold cross-validation of 79.00% ± 7.51 on the custom dataset and 71.15% ± 5.68 on SVD. The second-best results were obtained from the "big" 1D-CNN model fed with GTCC (78.83% ± 4.85 for the custom dataset and 71.18% ± 5.68 for SVD), followed by

the BiLSTMs fed with GTCCs (74.39% ± 6.15 for the custom dataset). However, due to the lack of consistent performance of the BiLSTMs on SVD (65.16% ± 3.60), the model was not selected for multimodal integration. The most stable and high-performing choice across both datasets was therefore GTCCs on the “small” 1D-CNN, which was adopted as the default audio configuration across all fusion strategies.

The results for laryngeal bioimpedance modality were notably clearer in multi-class discrimination than in the binary pathology detection task, with speech consistently outperforming the sustained phonation across all classifiers and both datasets (chapter 8). On the custom dataset, the Gammatone spectrograms processed with the “small” 1D-CNNs yielded the highest average accuracy (74.21% ± 5.41), followed by raw WAVs fed into the “big” 1D-CNNs (73.73% ± 5.01), and GTCCs on either 1D-CNN model. For the last architecture, the performance gap between the speech and sustained phonation was the most prominent reaching a 24% difference; the speech signals achieved 72.90% ± 5.00 in the “small” 1D-CNN while sustained phonation reached 48.52% ± 7.45, and for the “big” 1D-CNN the speech achieved 72.05% ± 4.43, while sustained phonation remained at 48.59% ± 6.15. During the SVD testing, the GTCCs fed into the “small” 1D-CNN model emerged as the best-performing feature extraction method (66.81% ± 6.14), followed by WAV files processed using the “big” 1D-CNN architecture (64.04% ± 6.08) and Gammatone spectrograms fed into the “small” 1D-CNN (63.21% ± 5.93).

Based upon the above findings, the multimodal classification systems were developed using solely the continuous speech signals for both data modalities. The combination of GTCCs for audio with Gammatone spectrograms and WAV files for bioimpedance signals were used as feature representations, and the 1D-CNNs and 2D-CNNs were applied as the main classification architectures, alternating depending on the fusion strategy. The features were derived accordingly and processed following the methodology outlined in chapter 7 of

this work. The following subsections present the results obtained for the early, hybrid, and late fusion approaches in multimodal laryngeal pathology classification, with the capability of detecting cancerous and precancerous lesions with high accuracy. The models' performance is evaluated in terms of accuracy, precision, sensitivity, specificity, and F1 scores, with comparisons drawn against the best-performing unimodal baselines.

9.2.1 Early Fusion for Laryngeal Pathology Classification

The details of the early fusion methodology implemented in this research are available in section 7.1 of this thesis (*7.1. Early Fusion*).

The early fusion strategy for the multimodal laryngeal pathology classification involved concatenating modality-specific features prior to inputting them into one shared DL classification network. Based on the results of the unimodal experiments, the most effective feature combinations were selected for the multimodal integration: for audio signals, the GTCCs were chosen as the feature extraction method, while the laryngeal bioimpedance was processed in the form of Gammatone spectrograms. Derived matrices were subsequently concatenated at the feature level and fed into the unified "small" 1D-CNN architecture. This decision was motivated by the consistent strong performance of both feature types across the custom dataset and SVD. The following figure (Figure 9.16) shows the confusion matrices calculated for the applied early fusion strategy in the multi-class laryngeal pathology classification tested on the custom dataset (Figure 9.16A) and SVD (Figure 9.16B).

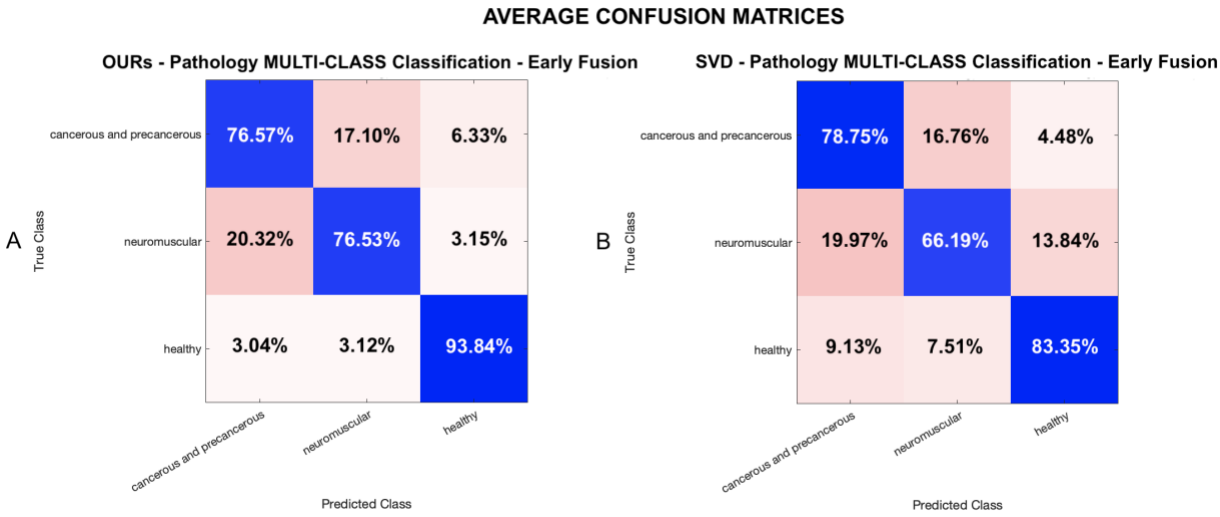

Figure 9.16: The average confusion matrices obtained for the designed early fusion multi-class classification multimodal system tested over 10-fold cross-validation on the custom dataset (figure A) and SVD (figure B).

According to the confusion matrices, the highest accuracy was achieved for the healthy cases, with both pathologies – cancerous and precancerous lesions, and neuromuscular disorders – slightly misclassified as adjacent pathologies. Otherwise, the early fusion strategy demonstrated solid and balanced classification performance, with the average accuracy calculated over the 10-fold cross-validation reaching 82.54% ± 4.61 on the custom dataset, and 76.10% ± 5.07 on SVD, with consistently high per-class accuracy: 84.69% ± 3.36 for cancerous and precancerous, 85.53% ± 3.06 for neuromuscular, and 94.85% ± 3.94 for healthy cases on the custom dataset, and for SVD: 83.22% ± 4.76 for cancerous and precancerous, 80.64% ± 4.55 for neuromuscular, and 88.34% ± 2.53 for healthy cases. Thus, the early fusion multimodal classification outperformed its unimodal counterparts on both datasets. Furthermore, while tested on the custom dataset, all performance metrics of the early fusion model achieved higher scores than those reached by the unimodal systems. In terms of precision, the model achieved 74.61% ± 4.22 (cancerous and precancerous), 81.34% ± 8.24 (neuromuscular), and 91.55% ± 6.20 (healthy). Sensitivity as well as F1 scores followed a similar trend, reaching respectively 76.57% ± 9.98 and 75.36% ± 6.03 for malignant cases, 76.53% ± 5.01 and 78.56% ± 4.11 for neuromuscular, and 93.84% ± 5.64

and 92.63% ± 5.53 for healthy classes. The following two figures represent the performance metrics calculated for the early fusion multi-class laryngeal pathology classification model tested on the custom dataset (Figure 9.17), and SVD (Figure 9.18).

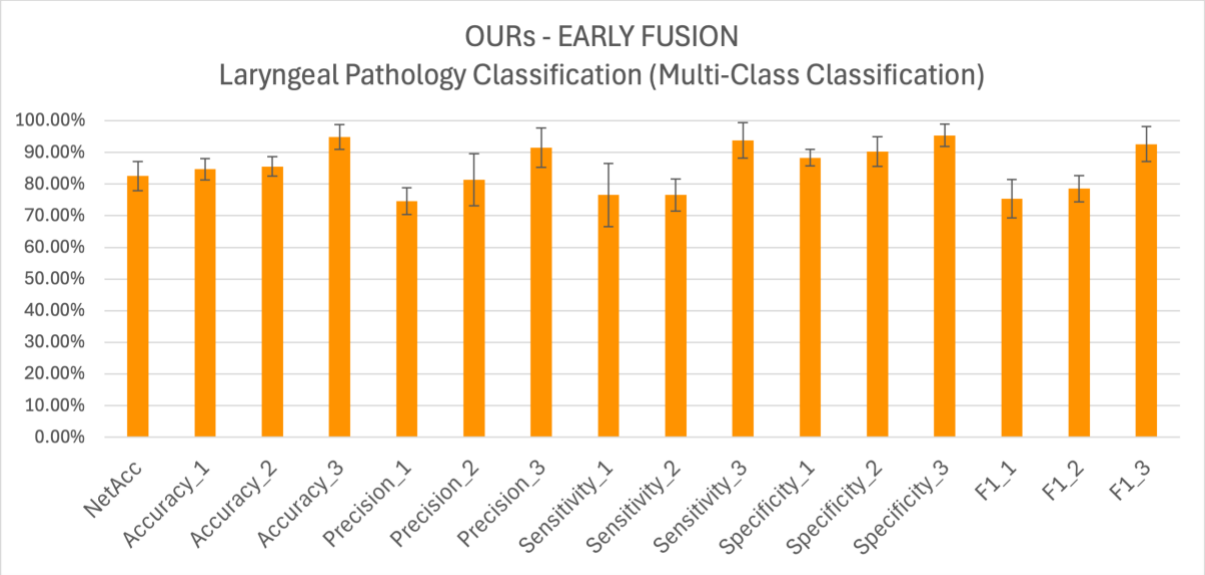

Figure 9.17: Early Fusion Model in Laryngeal Pathology Classification – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

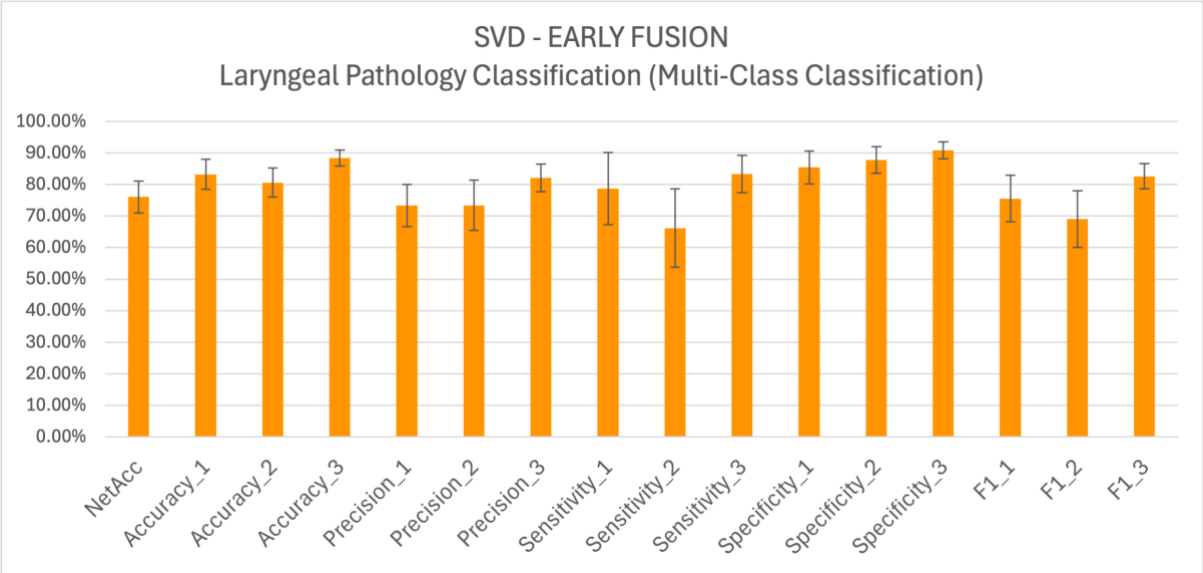

Figure 9.18: Early Fusion Model in Laryngeal Pathology Classification – the accuracy, precision, sensitivity, specificity and F1 scores calculated for SVD data testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

These results confirm that early fusion offers a balanced and interpretable classification framework. It presents a significant improvement over unimodal systems for both datasets tested, particularly in boosting classification reliability for cancerous and precancerous conditions, which is the primary aim of this study. Further comparisons between various multimodal approaches, as well as the direct comparison with the unimodal systems' metrics are further discussed in section 9.2.4 of this work.

9.2.2 Hybrid (Intermediate) Fusion for Laryngeal Pathology Classification

The details of the hybrid (intermediate) fusion methodology implemented in this research are available in section 7.2 of this thesis (*7.2. Hybrid (Intermediate) Fusion*).

The hybrid fusion strategy was designed to independently process each modality using a dedicated DL subnetwork (network branch), enabling modality-specific feature learning before the integration of the learned intermediate representations at the level of the concatenation layer. The hybrid fusion approach was motivated by the distinct temporal and spectral properties of the audio and bioimpedance signals (for instance, fusing the audio-derived GTCCs and raw EGG WAV files), and the aim of preserving and enhancing the modality-specific patterns before their combined evaluation and multimodal classification.

For the audio modality subnetwork, the input consisted of GTCCs processed using the 1D-CNN architecture. The audio 1D-CNN included two convolutional blocks with normalisation, ReLU activation, and 20% dropout layers, followed by a global average pooling and a fully connected layer followed by the flatten layer to prepare the feature vector for fusion (Figure 7.3).

For the laryngeal bioimpedance modality subnetwork, two separate feature extraction methods were explored: the raw waveform (WAV) data and the Gammatone spectrograms,

both derived from speech recordings. Each representation was processed using the 2D-CNN model – this choice was made due to the high performance of EGG data on 2D-CNNs in pathology detection (Table 8.16). The 2D-CNN branch consisted of four convolutional blocks with increasing depth, interleaved with normalisation, ReLU activation, and dropout, followed by global pooling, a fully connected layer, and a flattening layer (Figure 7.3).

The outputs from the audio and bioimpedance subnetworks were fused in the concatenation layer and passed through the two fully connected layers interleaved with a ReLU activation and a 20% rate dropout, following into the softmax function and the final classification layer. This configuration allowed the network to capture the modality-specific as well as the joint representations of the two modalities.

The following figure presents the average confusion matrices calculated over the 10-fold cross-validation of the two designed hybrid fusion multimodal multi-class classification systems tested on both datasets – the custom dataset (Figure 9.19A and 9.19B), and SVD (Figure 9.19C and 9.19D). The confusion matrices on the left side of the figure depict the raw EGG WAVs approach (Figure 9.19A and 9.19C), while those on the right – the EGG-derived Gammatone spectrograms (Figure 9.19B and 9.19D).

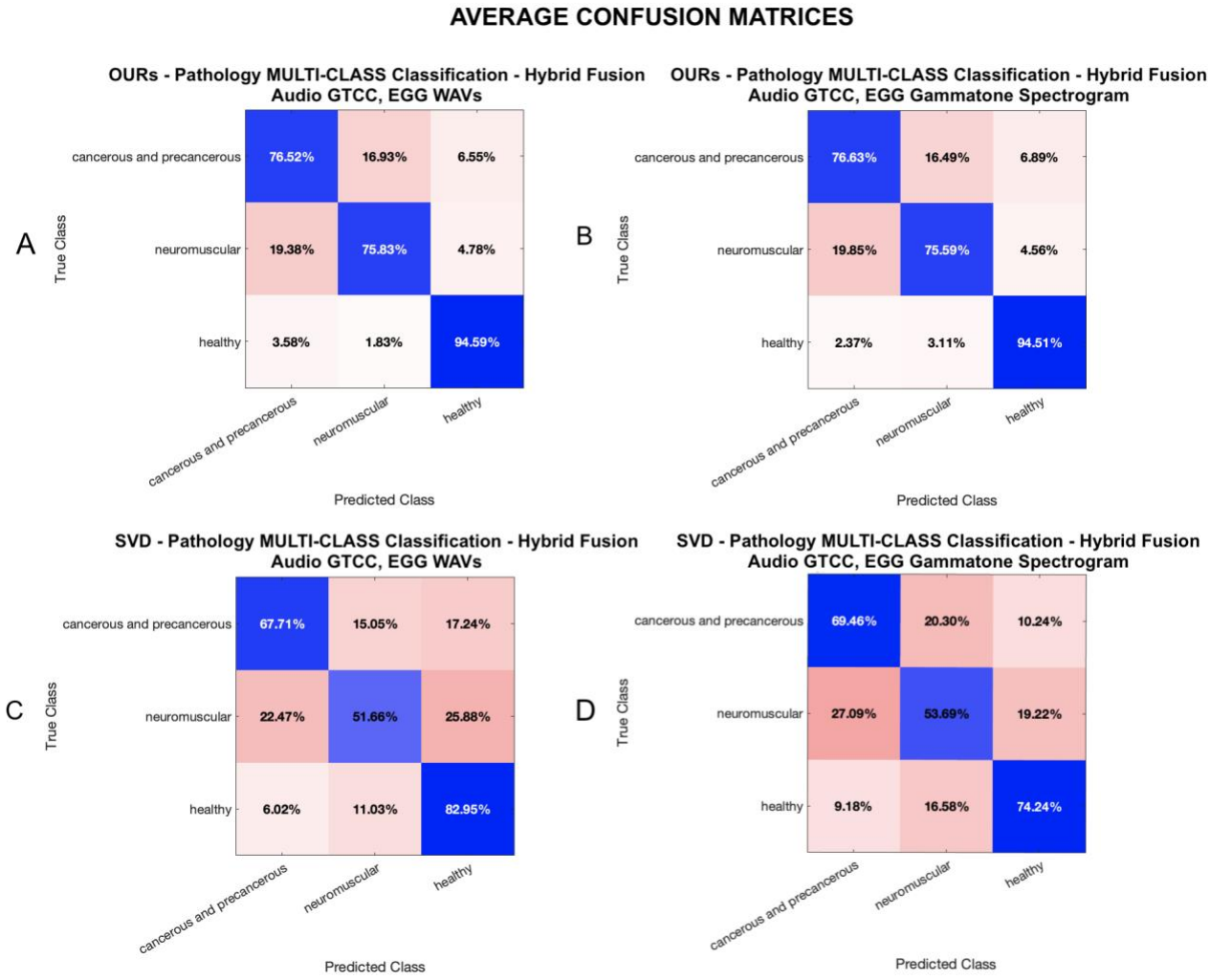

Figure 9.19: The average confusion matrices obtained for the designed hybrid fusion multimodal systems for laryngeal pathology classification, tested over 10-fold cross-validation on the custom dataset (figure A and B) and SVD (figure C and D). A and C present the confusion matrices calculated for EGG signals fed into the model as WAV files, and B and D show the confusion matrices calculated for EGG signals fed into the model as Gammatone spectrograms.

While tested on the custom dataset, the hybrid configurations yielded strong and relatively similar performance across all three pathology classes. However, unlike the hybrid model designed for the laryngeal pathology detection (the binary classification), the hybrid model designed for laryngeal pathology classification did not outperform its early fusion counterpart. The best hybrid model performance was obtained using raw WAV EGG signals, achieving a net accuracy of 82.67% ± 5.00%, with class-specific accuracies of 84.9% ± 3.65 (cancerous and precancerous), 85.95% ± 5.03 (neuromuscular), and 94.48% ± 3.15 (healthy). The precision of cancerous and precancerous lesions' identification increased

slightly over that achieved by the early fusion multimodal system, reaching 75.74% ± 7.82 using the EGG WAVs, and 76.86% ± 6.87 using EGG-derived Gammatone spectrograms. The other performance metrics remained comparable to those achieved using early fusion. Overall, while the hybrid models outperformed the unimodal laryngeal pathology classification systems, their performance on the custom dataset was comparable to that of the early fusion.

Nevertheless, the SVD testing of the designed multi-class hybrid fusion classification models delivered notably lower results, approaching those achieved by the audio-based unimodal classification system. Both the WAV- and Gammatone spectrogram-based hybrid configurations failed to match the accuracy achieved by the custom dataset, with reduced precision, sensitivity, and F1-scores observed across all three classes (Table 9.3 and 9.4). This performance gap is likely attributable to limitations inherent in the SVD dataset, which were previously outlined in section 4.4.1., particularly, the limited representation of certain pathologies, especially malignant cases, and the lack of balance across the classes (the number of data samples). As a result, while the hybrid systems generalised well within a controlled dataset, their performance dropped when applied to less consistent data. This outcome further supports the importance of using carefully curated datasets for developing and validating clinically applicable classification models.

The following figures visualise the performance metrics calculated for the two designed hybrid fusion multimodal multi-class laryngeal pathology classifiers designed for the purposes of this study and tested using the custom dataset (Figure 9.20 and 9.21), and SVD (Figure 9.22 and 9.23). Figures 9.20 and 9.22 show the results of the hybrid model with the bioimpedance subnetwork fed with raw WAVs of EGG signals, while Figures 9.21 and 9.23 depict the performance metrics obtained using EGG-derived Gammatone spectrograms as the input for the bioimpedance subnetwork of the hybrid fusion model.

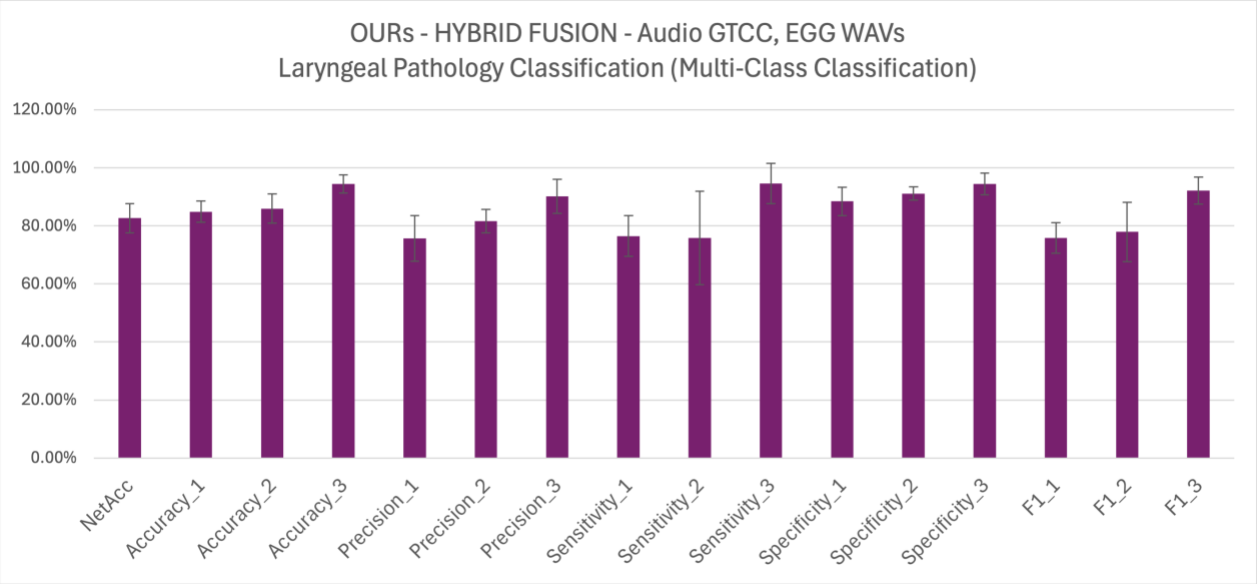

Figure 9.20: Hybrid Fusion Model in Laryngeal Pathology Classification fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

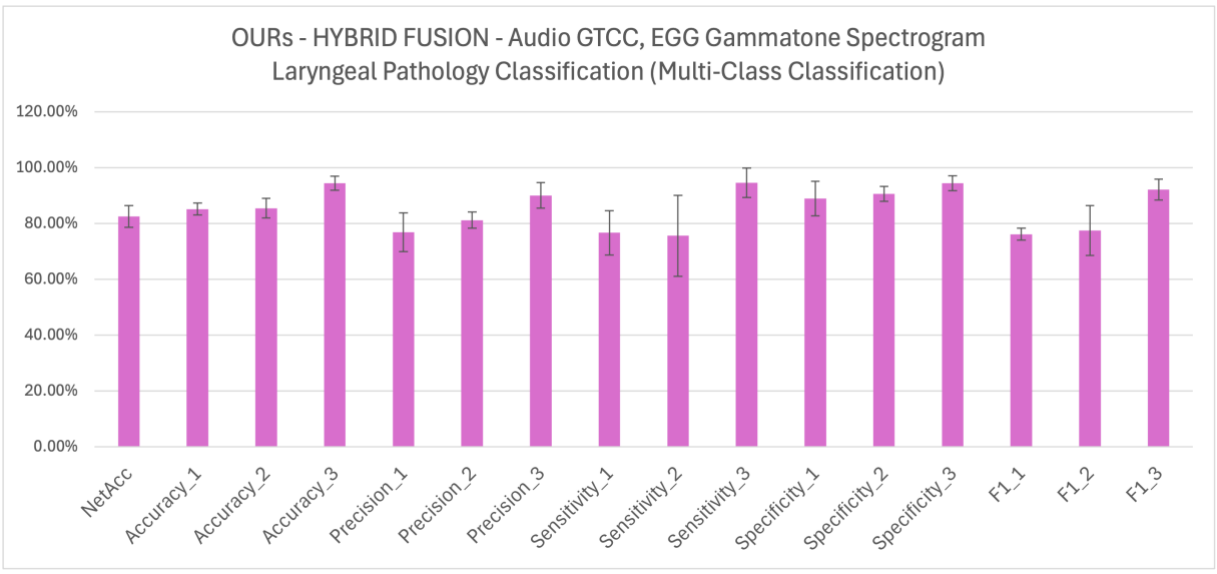

Figure 9.21: Hybrid Fusion Model in Laryngeal Pathology Classification fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

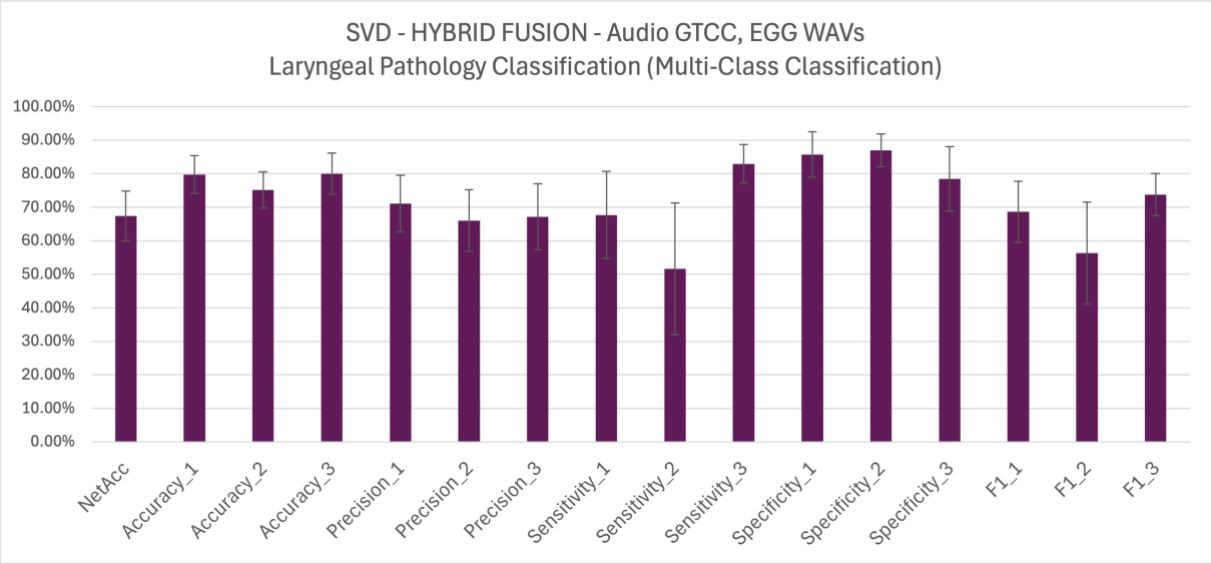

Figure 9.22: Hybrid Fusion Model in Laryngeal Pathology Classification fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

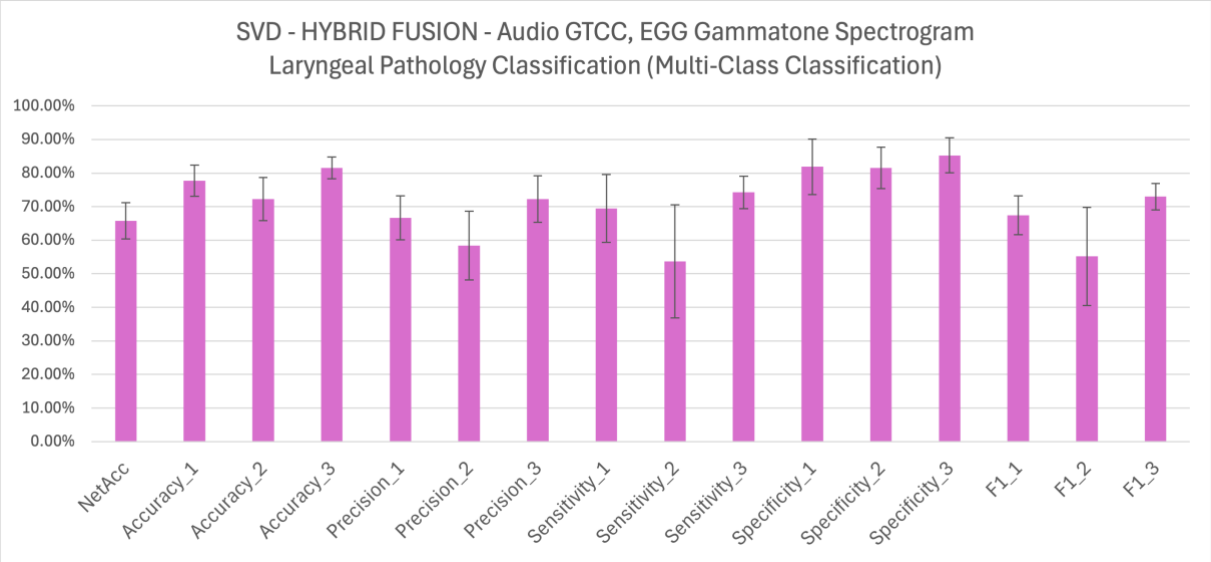

Figure 9.23: Hybrid Fusion Model in Laryngeal Pathology Classification fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

Overall, when tested with the custom dataset, the hybrid fusion strategy proved effective for multi-class classification, outperforming the unimodal systems and reaching similar

performance to that of the early fusion multi-class classification model. Nevertheless, the SVD testing of the hybrid fusion strategy questioned the method's effectiveness for the multi-class classification of laryngeal pathologies.

9.2.3 Late Fusion for Laryngeal Pathology Classification

The details of the late fusion methodology implemented in this research are available in section 7.3 of this thesis (*7.3. Late Fusion*).

The late fusion strategy employed in this study followed a stacked generalisation framework, where two individual modality-specific classifiers were trained independently, and their output predictions were later combined by a meta-classifier at the decision level. This approach offered the flexibility to tailor each branch to the strengths of its corresponding modality while avoiding early or intermediate constraints on feature compatibility.

Two late fusion configurations were examined:

1. Audio speech-derived GTCCs processed using the "small" 1D-CNN, combined with raw bioimpedance speech WAVs processed using the "big" 1D-CNN.
2. Audio speech-derived GTCCs processed using the "small" 1D-CNN, combined with bioimpedance speech-derived Gammatone spectrograms processed using the "small" 1D-CNN.

These configurations were selected based on the best-performing unimodal systems, discussed previously in chapter 8. Given the strong unimodal classification performance of the bioimpedance signals in a form of Gammatone spectrograms and raw WAV files, the evaluation of both configurations allowed for a direct comparison between spectrogram-based and waveform-based bioimpedance representations under the same late fusion strategy.

In both configurations, the class probability outputs of the two independently trained and validated classifiers (the modality-specific CNNs) were treated as the input features for the subsequent meta-classifier. Thus, the meta-classification system relied on the stacked generalisation technique applied using ECOC, learning the optimal decision boundaries by analysing cross-modal patterns in the prediction scores.

The following figure (Figure 9.24) presents the confusion matrices calculated for the late fusion multimodal multi-class laryngeal pathology classification systems designed for the purposes of this study, calculated over the 10-fold cross-validation approach. The confusion matrices found on the left side of the figure (Figure 9.24A and 9.24C) relate to the late fusion model with the bioimpedance-specific classifier fed with raw EGG signals in a WAV format, while those on the right (Figure 9.24B and 9.24D) present the confusion matrices obtained from the late fusion model where the bioimpedance-specific classifier was fed with the EGG-derived Gammatone spectrograms.

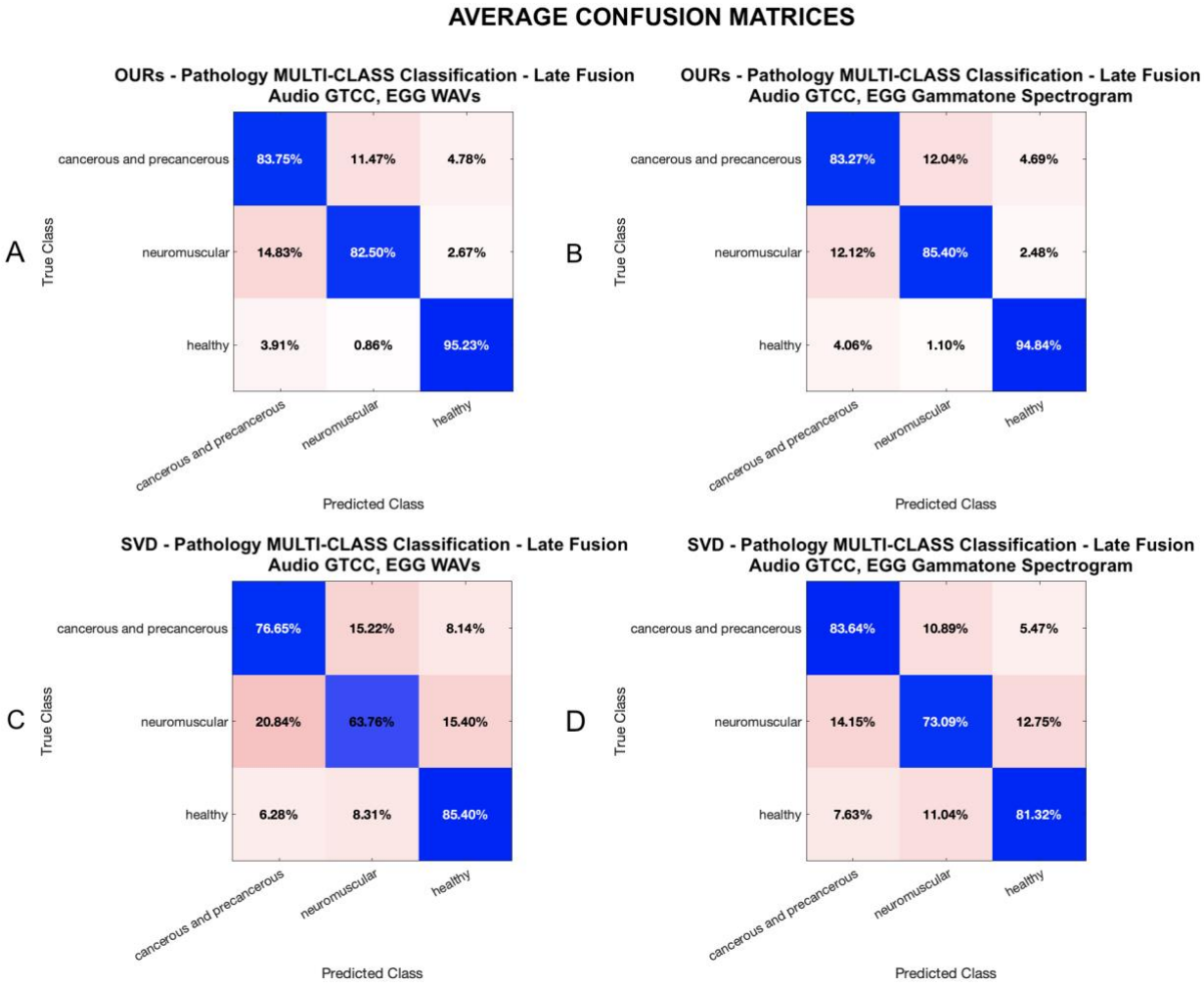

Figure 9.24: The average confusion matrices obtained for the designed late fusion multimodal systems for laryngeal pathology classification, tested over 10-fold cross-validation on the custom dataset (figure A and B) and SVD (figure C and D). A and C present the confusion matrices calculated for EGG signals fed into the model as WAV files, and B and D show the confusion matrices calculated for EGG signals fed into the model as Gammatone spectrograms.

The first configuration evaluated using the late fusion strategy – with the bioimpedance signals fed into the "small" 1D-CNN model as Gammatone spectrograms – outperformed the raw WAV file counterpart: the first configuration (raw EGG WAVs processed using the "big" 1D-CNN) achieved the overall classification accuracy of 87.27% ± 4.52 on the custom dataset, and 75.27% ± 4.78 on SVD, while the second configuration (EGG-Gammatone spectrograms processed using the "small" 1D-CNN) achieved 88.00% ± 2.29 on the custom dataset, and 79.35% ± 2.63 on SVD. On the custom dataset, the second late fusion model achieved per-class accuracies of 89.23% ± 1.95 for cancerous and precancerous lesions,

90.84% ± 3.00 for neuromuscular disorders, and 95.94% ± 2.02 for healthy cases. These results are the highest multi-class laryngeal pathology classification scores achieved by any model developed and tested during the course this research.

The designed late fusion model relying on the bioimpedance input as the EGG-derived Gammatone spectrograms processed using the “small” 1D-CNN architecture delivered the highest performance metrics among all designed multi-class classifiers – the precision reached 82.32% ± 2.78 for the malignant cases, 87.58% ± 3.84 for neuromuscular, and 93.5% ± 3.48 for healthy cases on the custom dataset, while for SVD the precision values for these classes were 79.85% ± 6.62, 77.62% ± 7.20, and 81.73% ± 4.74, respectively. Specificity achieved by this model, particularly for the malignant cases (91.90% ± 1.75 on the custom data, and 89.11% ± 4.52 on SVD), was also the highest of all developed and tested models. Providing the sensitivity and F1 scores achieved for the cancerous and precancerous cases (sensitivity and F1 of 83.27% ± 3.43 and 82.76% ± 2.60 on the custom dataset, and 83.64% ± 7.10 and 81.43% ± 4.76 on SVD, respectively), this model delivered the best performance for the accurate identification of malignancies.

The WAV-based configuration – although still outperforming the unimodal, early fusion, and hybrid fusion models for the custom dataset – performed slightly below its EGG-Gammatone spectrogram counterpart, with marginally lower accuracy across all classes. The following figures contribute the visual representation of the performance metrics calculated over 10-fold cross-validation for all late fusion multimodal laryngeal pathology classification models tested on the custom dataset (Figure 9.25 and 9.26) and SVD (Figure 9.27 and 9.28). Figures 9.25 and 9.27 show the results of the late fusion model with the bioimpedance classifier of the “big” 1D-CNN architecture fed with raw WAVs of EGG signals, while Figures 9.26 and 9.28 present the performance metrics obtained using EGG-derived

Gammatone spectrograms processed with the “small” 1D-CNN for the bioimpedance classifier.

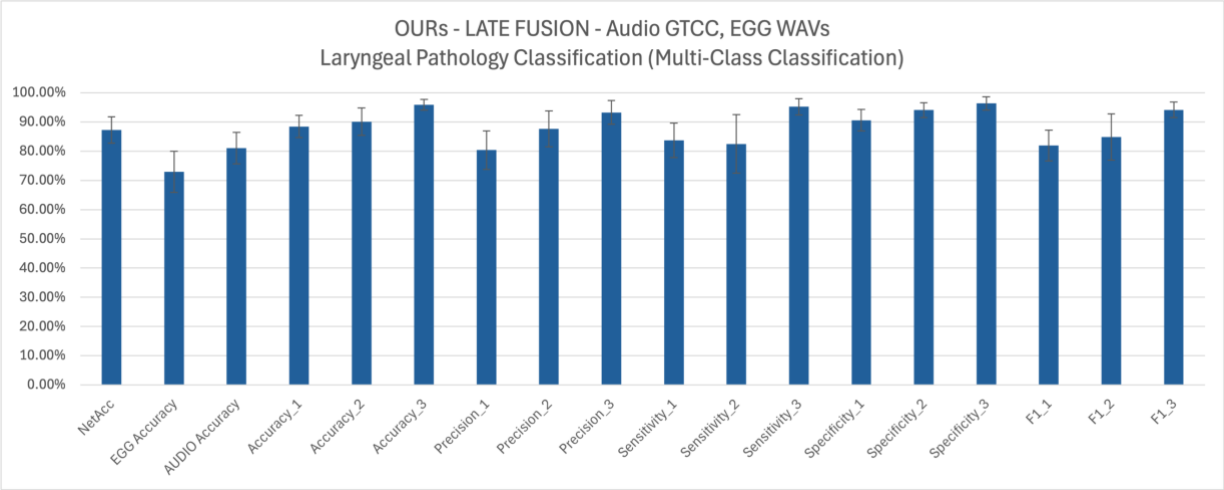

Figure 9.25: Late Fusion Model in Laryngeal Pathology Classification with bioimpedance classifier based on the “big” 1D-CNN architecture fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

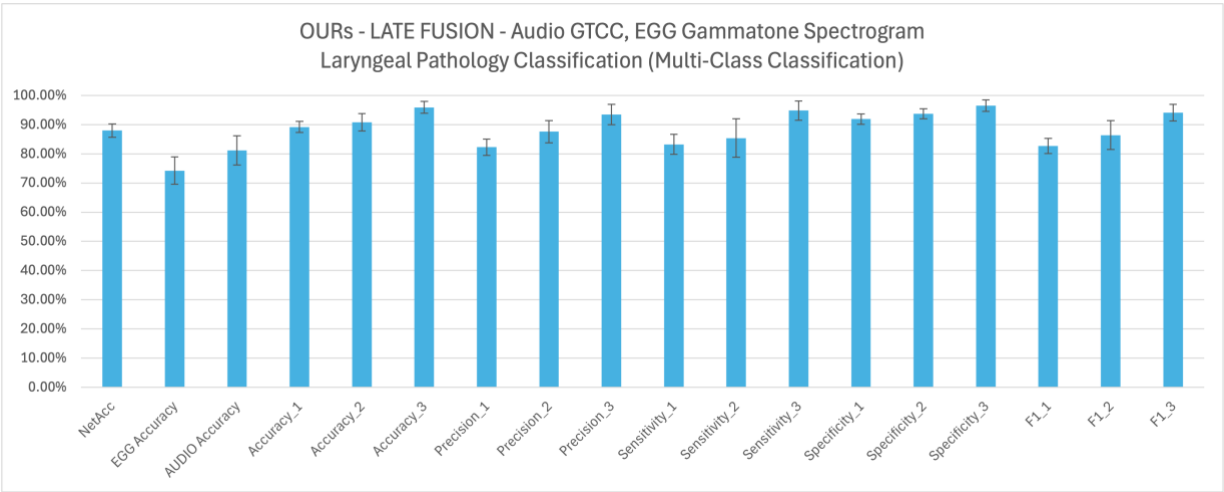

Figure 9.26: Late Fusion Model in Laryngeal Pathology Classification with bioimpedance classifier based on the “small” 1D-CNN architecture fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the custom dataset testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

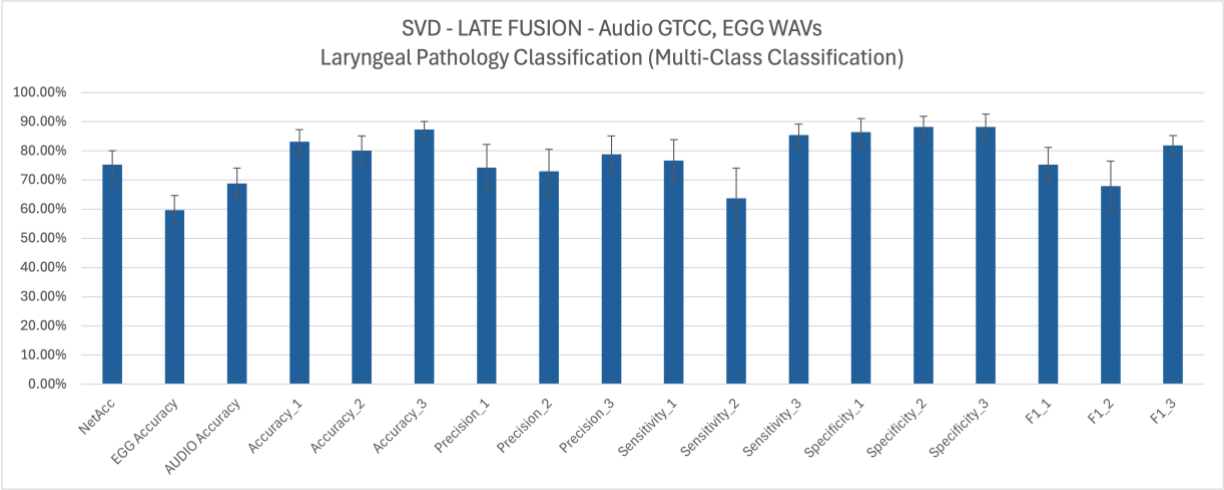

Figure 9.27: Late Fusion Model in Laryngeal Pathology Classification with bioimpedance classifier based on the "big" 1D-CNN architecture fed with EGG signals as WAVs – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

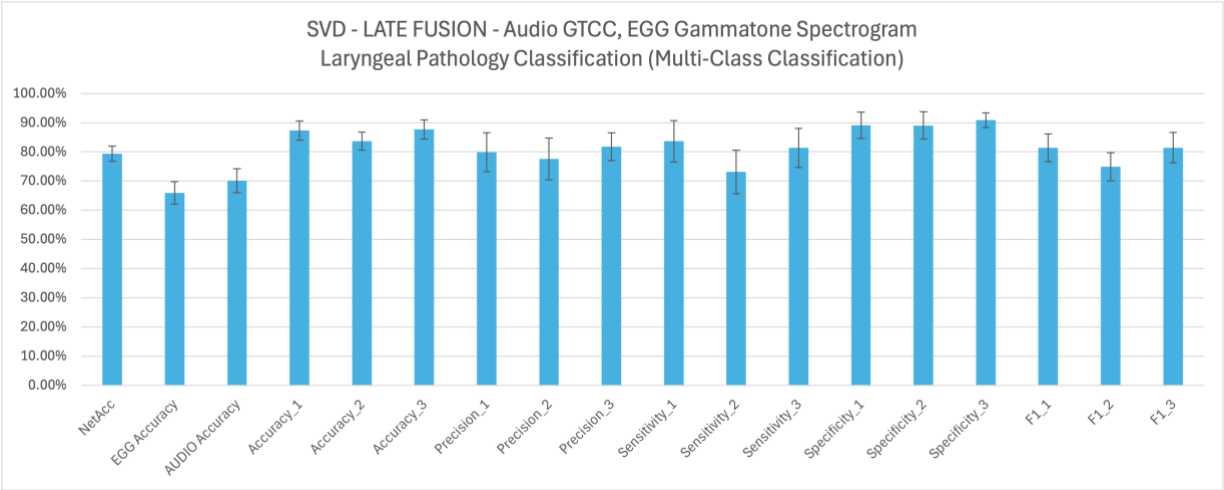

Figure 9.28: Late Fusion Model in Laryngeal Pathology Classification with bioimpedance classifier based on the "small" 1D-CNN architecture fed with EGG signals as Gammatone spectrograms – the accuracy, precision, sensitivity, specificity and F1 scores calculated for the SVD testing. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

These results suggest that while both modality configurations benefited from the late fusion architecture, the spectrogram-based model was better-suited for capturing and preserving the discriminative characteristics of bioimpedance signals, particularly for cancerous and precancerous cases.

In summary, late fusion provided the highest classification accuracy and the most balanced class-wise performance of all tested fusion strategies. The precision and sensitivity metrics achieved for the cancerous and precancerous cases proved the late fusion model based on EGG-derived Gammatone spectrograms as the input for the bioimpedance-specific classification branch delivers the best performance in term of the accurate identification of malignancies. These findings reinforce the value of stacked generalisation for multi-class clinical applications, making the audio-derived GTCCs combined with EGG-derived Gammatone spectrograms late fusion model not only the most accurate but also the most clinically promising system developed in this study for multi-class laryngeal pathology classification.

9.2.4 Conclusions on Multimodal Laryngeal Pathology Classification

This section presented and discussed the performance of the three multimodal fusion strategies – early, hybrid, and late – for multi-class classification of laryngeal pathologies using speech-based audio and laryngeal bioimpedance (EGG) signals. Across all architectures, multimodal systems outperformed their unimodal counterparts, reaffirming the advantage of integrating complementary physiological and acoustic modalities.

To conclude the multi-class laryngeal pathology classification analysis, a comparative evaluation of the best performing unimodal and all multimodal systems was conducted across the following performance metrics: overall model accuracy, as well as class-specific accuracy, precision, sensitivity, specificity, and F1-score. The following tables show the average performance metrics calculated over the 10-fold cross-validation performed on the best performing unimodal systems and all multimodal classifiers designed for the purposes of multi-class laryngeal pathology classification – Table 9.3 presents the results of the

custom dataset testing, while Table 9.4 depicts the values of performance metrics achieved on SVD. The group numbers have been assigned to the corresponding models as follows:

- 1: BEST UNIMODAL on AUDIO (for custom dataset: GTCC on “small” 1D-CNN; for SVD: GTCC on “big” 1D-CNN),
- 2: BEST UNIMODAL on EGG (for custom dataset: Gammatone spectrograms on “small” 1D-CNN; for SVD: GTCC on “small” 1D-CNN),
- 3: EARLY FUSION (audio GTCC and EGG Gammatone spectrograms on “small” 1D-CNN),
- 4: HYBRID (audio GTCC on “small” 1D-CNN, EGG WAVs on 2D-CNN),
- 5: HYBRID (audio GTCC on “small” 1D-CNN, EGG Gammatone Spectrograms on 2D-CNN),
- 6: LATE FUSION STACKED (audio GTCC on “small” 1D-CNN, EGG WAVs on “big” 1D-CNN),
- 7: LATE FUSION STACKED (audio GTCC on “small” 1D-CNN, EGG Gammatone Spectrograms on “small” 1D-CNN).

Table 9.3: Performance metrics calculated for the best performing unimodal systems and all designed multimodal systems designed for the purposes of the laryngeal pathology classification (multi-class classification) tested using the custom dataset. “CA” stands for classification accuracy. The parameters were numbered according to the class, where 1 stands for cancerous and precancerous, 2 stands for neuromuscular, 3 stand for healthy. “A” is accuracy for the particular class, “P” is precision, “Sn” is sensitivity, “Sp” is specificity, and “F1” is the F1 score.

OUR-CM	CA	A1	A2	A3	P1	P2	P3	Sn1	Sn2	Sn3	Sp1	Sp2	Sp3	F1-1	F1-2	F1-3
1	79.00% ± 7.51	81.61% ± 6.38	82.38% ± 6.00	94.00% ± 3.57	71.95% ± 11.64	75.69% ± 10.15	89.49% ± 4.22	69.53% ± 8.92	72.85% ± 11.93	93.42% ± 8.84	86.96% ± 7.55	87.35% ± 5.56	94.28% ± 2.21	70.26% ± 8.29	73.87% ± 9.36	91.26% ± 5.71
2	74.21% ± 5.41	78.35% ± 5.37	83.15% ± 5.69	86.93% ± 6.39	66.38% ± 8.34	77.31% ± 9.17	80.63% ± 11.21	64.04% ± 12.20	74.54% ± 15.31	83.20% ± 14.64	84.87% ± 5.85	87.66% ± 7.69	88.89% ± 7.27	64.54% ± 7.96	74.76% ± 10.12	81.12% ± 10.24
3	82.54% ± 4.61	84.69% ± 3.36	85.53% ± 3.06	94.85% ± 3.94	74.61% ± 4.22	81.34% ± 8.24	91.55% ± 6.2	76.57% ± 9.98	76.53% ± 5.01	93.84% ± 5.64	88.31% ± 2.63	90.26% ± 4.7	95.39% ± 3.54	75.36% ± 6.03	78.56% ± 4.11	92.63% ± 5.53

4	82.67% ± 5	84.9 % ± 3.65	85.95% ± 5.03	94.48% ± 3.15	75.74% ± 7.82	81.7 % ± 4.05	90.21% ± 5.81	76.52% ± 6.97	75.83% ± 16.06	94.59% ± 6.94	88.51% ± 4.89	91.13% ± 2.28	94.46% ± 3.68	75.84% ± 5.26	77.97% ± 10.22	92.14% ± 4.66
5	82.53% ± 3.93	85.14% ± 2.16	85.47% ± 3.48	94.44% ± 2.56	76.86% ± 6.87	81.19% ± 2.87	90% ± 4.57	76.63% ± 7.9	75.59% ± 14.5	94.51% ± 5.25	88.88% ± 6.17	90.61% ± 2.65	94.42% ± 2.73	76.15% ± 2.1	77.42% ± 8.94	92.09% ± 3.73
6	87.27% ± 4.52	88.46% ± 3.79	90.12% ± 4.72	95.95% ± 1.84	80.38% ± 6.64	87.7 % ± 6.18	93.28% ± 4.06	83.75% ± 5.9	82.5 % ± 9.99	95.23% ± 2.77	90.62% ± 3.65	94.09% ± 2.54	96.37% ± 2.26	81.91% ± 5.26	84.92% ± 7.93	94.19% ± 2.64
7	88% ± 2.29	89.23% ± 1.95	90.84% ± 3	95.94% ± 2.02	82.32% ± 2.78	87.58% ± 3.84	93.5 % ± 3.48	83.27% ± 3.43	85.4 % ± 6.58	94.84% ± 3.27	91.9 % ± 1.75	93.69% ± 1.72	96.52% ± 1.96	82.76% ± 2.6	86.42% ± 4.96	94.14% ± 2.89

Table 9.4: Performance metrics calculated for the best performing unimodal systems and all designed multimodal systems designed for the purposes of the laryngeal pathology classification (multi-class classification) tested using SVD. "CA" stands for classification accuracy. The parameters were numbered according to the class, where 1 stands for cancerous and precancerous, 2 stands for neuromuscular, 3 stand for healthy. "A" is accuracy for the particular class, "P" is precision, "Sn" is sensitivity, "Sp" is specificity, and "F1" is the F1 score.

SVD-CM	CA	A1	A2	A3	P1	P2	P3	Sn1	Sn2	Sn3	Sp1	Sp2	Sp3	F1-1	F1-2	F1-3
1	71.18% ± 5.68	81.12% ± 4.47	77.27% ± 5.37	83.96% ± 4.25	72.48% ± 8.84	65.71% ± 7.53	75.8 % ± 6.19	71.28% ± 4.81	65.84% ± 12.39	76.41% ± 10.11	86.04% ± 5.25	82.99% ± 4	87.74% ± 3.6	71.72% ± 5.97	65.54% ± 9.43	75.86% ± 7.12
2	66.81% ± 6.14	74.1 % ± 6.62	75.65% ± 2.58	83.88% ± 4.24	58.8 % ± 7.65	74.19% ± 6.67	75.52% ± 7.28	79.98% ± 12.44	43.11% ± 13.84	77.34% ± 6.4	71.16% ± 9.38	91.91% ± 4.89	87.14% ± 5	67.26% ± 7.67	53.02% ± 9.95	76.24% ± 5.7
3	76.1 % ± 5.07	83.22% ± 4.76	80.64% ± 4.55	88.34% ± 2.53	73.34% ± 6.69	73.4 % ± 8.01	82.13% ± 4.34	78.75% ± 11.45	66.19% ± 12.47	83.35% ± 5.9	85.45% ± 5.23	87.86% ± 4.23	90.84% ± 2.64	75.58% ± 7.41	69.08% ± 8.97	82.62% ± 3.98
4	67.44% ± 7.45	79.74% ± 5.61	75.19% ± 5.42	79.95% ± 6.17	71.12% ± 8.49	66.02% ± 9.26	67.14% ± 9.81	67.71% ± 12.94	51.66% ± 19.61	82.95% ± 5.74	85.76% ± 6.75	86.96% ± 4.88	78.44% ± 9.63	68.7 % ± 9.07	56.33% ± 15.21	73.76% ± 6.27
5	65.8 % ± 5.43	77.73% ± 4.61	72.27% ± 6.4	81.59% ± 3.23	66.66% ± 6.57	58.43% ± 10.22	72.25% ± 6.92	69.46% ± 10.15	53.69% ± 16.86	74.24% ± 4.84	81.87% ± 8.26	81.56% ± 6.16	85.27% ± 5.21	67.45% ± 5.76	55.17% ± 14.66	72.97% ± 3.93
6	75.27% ± 4.78	83.17% ± 4.14	80.08% ± 4.98	87.29% ± 2.77	74.31% ± 7.9	72.99% ± 7.6	78.88% ± 6.22	76.65% ± 7.23	63.76% ± 10.3	85.4 % ± 3.75	86.44% ± 4.6	88.23% ± 3.64	88.23% ± 4.33	75.25% ± 5.96	67.87% ± 8.58	81.84% ± 3.38
7	79.35% ± 2.63	87.29% ± 3.26	83.72% ± 3.1	87.7 % ± 3.28	79.85% ± 6.62	77.62% ± 7.2	81.73% ± 4.74	83.64% ± 7.1	73.09% ± 7.47	81.32% ± 6.77	89.11% ± 4.52	89.03% ± 4.68	90.89% ± 2.55	81.43% ± 4.76	74.91% ± 4.84	81.44% ± 5.19

These results are also summarised in the following two figures, which visually compare all performance parameters calculated for the models – Figure 9.29 shows the results obtained for the custom dataset testing, while Figure 9.30 depicts the results calculated for the SVD testing.

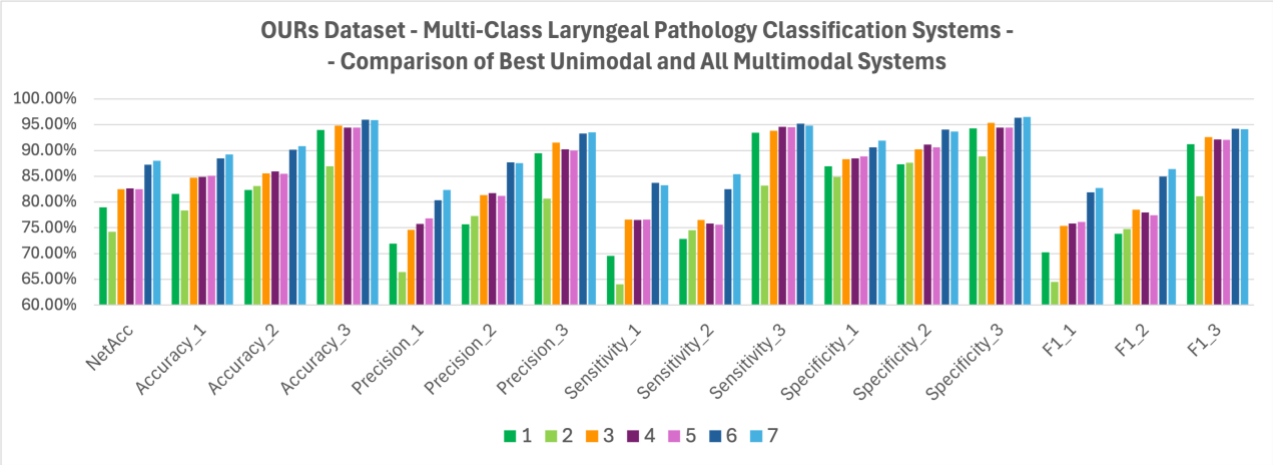

Figure 9.29: Visual representation of the results obtained for the multi-class laryngeal pathology classification models trained and validated on the custom dataset, including the best performing unimodal laryngeal pathology classification systems and all multimodal systems designed, depicting the overall model's accuracy ("NetAcc"), as well as class-specific accuracy, precision, sensitivity, specificity, and F1 score parameters. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

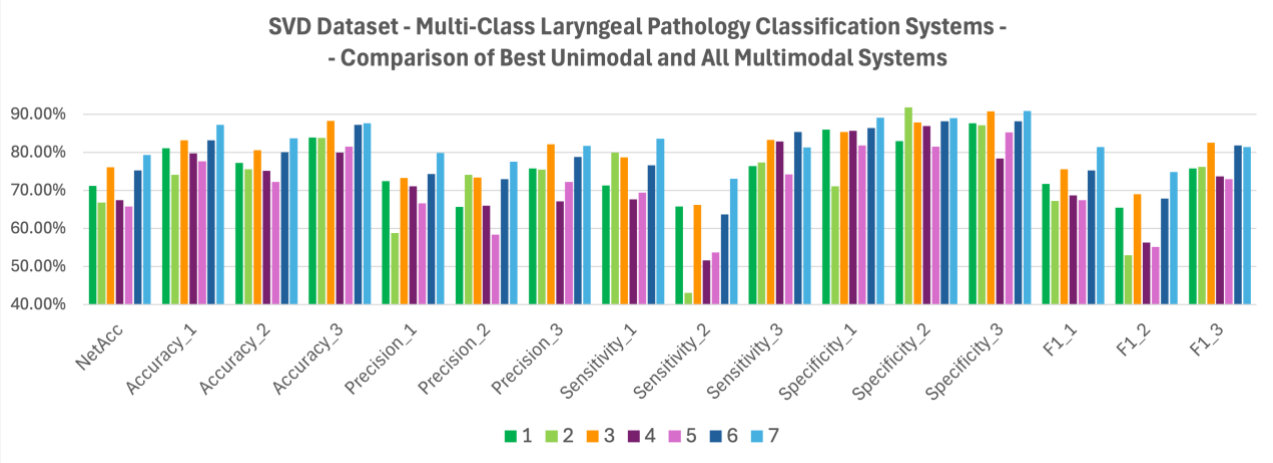

Figure 9.30: Visual representation of the results obtained for the multi-class laryngeal pathology classification models trained and validated on SVD, including the best performing unimodal laryngeal pathology classification systems and all multimodal systems designed, depicting the overall model's accuracy ("NetAcc"), as well as class-specific accuracy, precision, sensitivity, specificity, and F1 score parameters. The parameters were numbered according to the class: 1 – cancerous and precancerous, 2 – neuromuscular, 3 – healthy.

According to the performance metrics calculated over the 10-fold cross-validation performed on all multimodal systems and best-performing unimodal systems, the most significant performance gains were observed for the late fusion models, where the two modalities – the audio recordings and the laryngeal bioimpedance signals – were fused at the decision level using ECOC meta-classifier. Notably, the best-performing system overall was the late fusion stacking model relying on audio-derived GTCCs, combined with the EGG-derived Gammatone spectrograms, both processed using the designed “small” 1D-CNN architecture. This model achieved the highest overall accuracy (88.00% on the custom dataset and 79.35% on SVD) and demonstrated the most balanced class-wise performance, with the highest precision and F1 scores achieved for cancerous and precancerous lesions across all the developed classifiers (respectively: 82.32% and 82.76% for the custom dataset, and 79.85% and 81.43% for SVD). These results were particularly important, since the primary goal of this study was the development of a multi-class laryngeal pathology classification system capable of identifying the malignancies with the highest accuracy. The designed late fusion model was also the only system to maintain superior classification accuracy when applied to SVD public dataset, underscoring its generalisability and practical reliability.

The early fusion model, which combined audio-derived GTCCs and bioimpedance-derived Gammatone spectrograms (through concatenation of the feature matrices) and processed them through a shared 1D-CNN, demonstrated strong baseline performance. With an overall accuracy of 82.54% on the custom dataset, and 76.10% on SVD, and well-balanced class-wise metrics, it served as an efficient and compact solution. However, the model’s inability to separate modality-specific representations slightly limited its discriminative power compared to more complex fusion approaches.

The hybrid fusion systems, which used separate DL branches for each modality prior to the network fusion (and learnt feature concatenation), offered improved class-wise performance and greater flexibility in feature representation. Both the WAV- and Gammatone-based EGG branches, when paired with GTCC audio input, performed comparably, with slight advantages observed for the WAV-based system. However, the improvements were modest, and neither hybrid model achieved dominance across all metrics. Furthermore, the hybrid models significantly underperformed during the SVD testing.

To complement the descriptive performance metrics discussed above, one-way ANOVA followed by Tukey's HSD post-hoc comparisons was conducted on the best-performing unimodal systems (audio and EGG) and all multimodal fusion strategies (early, hybrid, and late fusion) for both datasets.

For the custom dataset, ANOVA revealed a highly significant main effect of model type on classification accuracy, $F(6,63) = 8.44$, $p < 0.000001$, with a large effect size (partial eta squared) of $\eta_p^2 = 0.446$. Tukey's HSD comparisons demonstrated that the late fusion models (models 6 and 7) significantly outperformed both unimodal baselines, with all $p < 0.007$. Specifically, the EGG-only system (model 2) performed significantly worse than both late fusion approaches (mean difference -0.101, 95% CI $= [-0.157, -0.045]$, $p < 0.00001$ for model 2 and 6, and mean difference -0.109, 95% CI $= [-0.164, -0.053]$, $p < 0.000003$ for model 2 and 7). Likewise, the audio-only model (model 1) significantly underperformed in relation to the late fusion classifiers (mean difference -0.068, 95% CI $= [-0.124, -0.013]$, $p = 0.007$ for model 1 and 6, and mean difference -0.076, 95% CI $= [-0.132, -0.020]$, $p = 0.002$ for model 1 and 7). No significant differences were observed between the two late fusion systems (models 6 vs 7: $p = 0.9995$), indicating that both variants provided comparably high performance.

For SVD, the results were consistent and followed the same trends as the custom dataset. ANOVA indicated a highly significant effect of model type ($F(6,63) = 7.62$, $p < 0.000004$), with a large effect size – partial eta squared $\eta_p^2 = 0.420$. Post-hoc analysis confirmed that unimodal EGG (model 2) performed significantly worse than the late fusion models (mean difference -0.132, 95% CI $= [-0.198, -0.065]$, $p < 0.000002$ for model 2 and 7). The audio-only baseline (model 1) also showed improvements over EGG (mean difference 0.070, 95% CI $= [0.004, 0.137]$, $p < 0.032$), but it did not outperform most multimodal systems. Consistent with the custom dataset, the two late fusion approaches (models 6 and 7) achieved the strongest overall performance, with model 7 significantly outperforming several alternative fusion designs (e.g., model 4: mean difference -0.075, 95% CI $= [-0.142, -0.009]$, $p = 0.016$; or model 5: mean difference -0.096, 95% CI $= [-0.162, -0.029]$, $p = 0.001$).

These results demonstrate that multimodal integration consistently provided significant performance gains over unimodal systems, and that late fusion strategies were the most effective in both datasets. Importantly, the improvements were not only statistically significant but also associated with large effect sizes, underscoring the robustness and practical relevance of the multimodal fusion for multi-class laryngeal pathology classification.

In summary, the results from this chapter confirm that the late fusion using stacked generalisation offers the most robust and scalable solution for multi-class laryngeal pathology classification, capable of detecting cancerous and precancerous lesions with the highest precision and sensitivity. It successfully harnesses the unique strengths of both modalities while mitigating their limitations through strategic class probability integration using a meta-classifier. The findings from this section not only validate the proposed multimodal framework but also position the late fusion model as the most clinically viable system developed in this study.

Chapter 10

Conclusions

This thesis presented a comprehensive investigation into the detection and classification of laryngeal pathologies using a multimodal deep learning approach based on audio recordings and simultaneous laryngeal bioimpedance measurements (gathered using electroglottography) collected from the control subjects with no diagnosis of a laryngeal pathology present, as well as participants suffering from various laryngeal conditions. Those included malignant lesions of the vocal folds, other growths not affecting vocal folds, neuromuscular disorders such as vocal fold paralysis, laryngitis, Reinke's Oedema, and lastly, functional dysphonia. To develop the intended multi-class laryngeal pathology classification system, we chose the following laryngeal conditions: cancerous and precancerous lesions, neuromuscular disorders, and control group of healthy participants.

The research was motivated by the critical need for non-invasive, efficient, and accurate preliminary diagnostic tools for laryngeal diseases, particularly for the early identification of malignancies. Following a systematic research rationale, the research hypothesis was formed, and a series of aims and objectives were set to guide the project's development. The original contributions to knowledge were outlined, together with the dissemination of findings through multiple scientific publications.

In this thesis, an in-depth review of the background knowledge necessary for the appropriate design and development of a laryngeal pathology classification system was first provided, introducing the anatomy of the human phonatory system and the nature of laryngeal pathologies, and detailing the physiological bases of the two data modalities chosen for the collection of the relevant data – the audio recordings of speech and sustained phonation, and the simultaneous electroglottographic measurements of laryngeal

bioimpedance. A review of the clinical profiles of the investigated laryngeal pathologies was also provided. We then provided a review of the available literature on the existing laryngeal pathology detection and classification systems that rely on the application of audio, EGG signals, or both modalities in a multimodal approach. Through this review, we exposed the present gaps in knowledge.

Subsequently, the creation and preprocessing of the custom dataset was documented. The development of the custom dataset is a major contribution to knowledge produced during this study. The publicly available SVD was also introduced as another dataset used for testing of the classification systems developed for the purposes of this study. Despite several critical limitations of the public dataset, SVD was used for completeness purposes, as well as to showcase the generalisability of the systems developed during this research. Initial data analysis explored the statistical properties and separability of the classes using statistical and time-frequency parameters and PCA. Overall, the data analysis results showed pathological and healthy signals can be separated to an extent based solely on the statistical methods, however, for more detailed multi-class laryngeal condition classification, more sophisticated methodologies, such as deep learning-based classification, are necessary.

The methodological chapters focused on the employed feature extraction methods, as well as the machine learning and deep learning architectures used in this study. The feature extraction methods were chosen based on the data analysis findings and included raw waveforms of the recorded signals (WAV files), STFT-based representations of the magnitude spectrum of those signals, as well as Mel-spectrum and ERB-spectrum based methods, including spectrograms and cepstral coefficients. Establishing that in laryngeal pathology detection and classification the ERB-based features outperform those considered

a “golden standard” in speech processing – the Mel-spectrum-based features – is another major contribution of this research.

The classification architectures included ensemble learning methods, CNNs (one-dimensional and two-dimensional), and RNNs (LSTM and BiLSTM). A dedicated chapter on multimodal fusion strategies – early, hybrid (intermediate), and late fusion – established the experimental design for combining modalities at different stages of processing.

For the purposes of accurate and robust laryngeal pathology detection and classification, two types of human phonation were investigated – sustained vowel phonation, and continuous speech. The testing performed on both phonation types during this study revealed that speech signals significantly outperform the sustained phonation in laryngeal pathology detection, as well as the multi-class classification of various conditions. This finding constitutes a novel contribution to knowledge, providing an insight contrary to most of the available literature (Rosa *et al.*, 1999; Jiang *et al.*, 1998; Henríquez *et al.*, 2009; Markaki and Stylianou, 2011; Islam *et al.*, 2022).

The experimental results chapters separately detailed unimodal and multimodal system performance for both binary detection (healthy vs pathological) and multi-class classification (malignant, neuromuscular, healthy). The results demonstrated the significant benefits of multimodal learning and fusion strategies, particularly the superior performance of late fusion methods. The systematic comparison of various multimodal data fusion methods in laryngeal pathology classification, as well as the results confirming that the integration of audio and laryngeal bioimpedance facilitates more accurate detection and classification of laryngeal pathologies compared to single-modality models constitute two major contributions to knowledge of this study.

This final chapter draws together the findings of the entire project. First, we provide a comprehensive summary and comparison of the results from the unimodal and multimodal

systems designed for the purposes of laryngeal pathology detection (binary classification) and multi-class laryngeal pathology classification capable of identifying the cancerous and precancerous lesions (section 10.1). In this section, we include the trends identified during the data analysis. Subsequently, the research hypothesis is revisited considering the findings and confirming its validity. The chapter then lists the contributions made to the fields of bio-medical digital signal processing and machine learning for laryngeal health. Finally, we outline potential directions for the future work, identifying the opportunities to extend and enhance the developed systems.

10. CONCLUSION

10.1. CONCLUDING REMARKS ON RESULTS AND THEIR COMPARISON

This study proposed and evaluated a series of unimodal and multimodal classification systems for the detection and classification of laryngeal pathologies using audio and laryngeal bioimpedance signals. The performance of each system was rigorously assessed across both binary (pathological vs healthy) and multi-class classification tasks (cancerous and precancerous, neuromuscular, healthy), supported by a comprehensive data analysis, with three multimodal fusion strategies (early, hybrid, and late) tested and compared. The testing performed led to the conclusions documented in the following.

10.1.1 Sustained Phonation versus Continuous Speech

Based on the results obtained during the examination of the unimodal systems, continuous speech consistently outperformed sustained phonation. For instance, the best-performing unimodal audio-based laryngeal pathology detection system ("small" 1D-CNN architecture fed with GTCC) achieved an average accuracy of 89.17% ± 7.30 when fed with the recordings of speech. Its sustained phonation counterpart achieved an average accuracy of 82.86% ± 4.29. The second best-performing audio-based pathology detection system reached 87.33% ± 11.07 on speech data, while its sustained phonation counterpart reached 82.79% ± 8.58.

However, in the EGG-based unimodal laryngeal pathology detection, the difference in performance between the speech and sustained phonation was less prominent – while in most of the best-performing models, speech outperformed the sustained phonation (for example: 2D-CNN on WAVs – speech achieving 86.03% ± 4.94 and sustained phonation reaching 82.96% ± 7.84; the "small" 1D-CNN on Gammatone spectrograms – speech achieving 84.83% ± 6.23 and sustained phonation reaching 82.99% ± 5.65; and the "big"

1D-CNN on Gammatone spectrograms – with speech achieving 84.56% ± 4.50 and sustained phonation reaching 82.36% ± 5.87), the model that delivered the highest accuracy across all developed EGG-based architectures was the 2D-CNN fed with STFT spectrograms derived from the sustained phonation – the system reached 87.39% ± 2.50 accuracy, while its speech-based counterpart reached 80.82% ± 3.82. Nevertheless, while tested on SVD, the EGG-based laryngeal pathology detection model delivered higher accuracy when fed with the speech signals than the sustained phonation (the best performing model: the "small" 1D-CNN on GTCC – speech achieving 80.99% ± 1.35 and sustained phonation reaching 72.59% ± 3.45).

All multi-class unimodal laryngeal pathology classification models (both audio-based as well as the EGG-based) delivered higher accuracy for speech signals than sustained phonation: the best-performing audio-speech-based laryngeal pathology classification model delivered the average accuracy of 79.00% ± 7.51 ("small" 1D-CNN on GTCC), while its sustained phonation counterpart reached 50.86% ± 6.39, resulting in nearly 30% difference in accuracy between the speech and sustained phonation in multi-class laryngeal pathology classification. ANOVA and Tukey's HSD confirmed the statistical significance of these results: $F(1,18) = 46.88$, $p = 0.000002$, mean difference of 0.197 with 95% CI = $[0.137, 0.258]$ and the effect size $\eta_p^2 = 0.723$ were achieved for the custom dataset; $F(1,18) = 21.88$, $p = 0.00019$, mean difference of 0.093 with 95% CI = $[0.051, 0.135]$, and the effect size $\eta_p^2 = 0.549$ were reported for SVD. Similarly, the best-performing EGG-speech-based laryngeal pathology classification model delivered an average accuracy of 74.21% ± 5.41 ("small" 1D-CNN on Gammatone spectrograms), while its sustained phonation counterpart reached 52.94% ± 5.49. Here, ANOVA and Tukey's HSD resulted in $F(1,18) = 28.09$, $p = 0.00005$, mean difference of 0.141 with 95% CI = $[0.084, 0.196]$, and the effect size $\eta_p^2 = 0.610$ for the custom dataset; $F(1,18) = 25.65$, $p = 0.00008$, mean

difference of 0.104 with 95% CI $= [0.061, 0.148]$, and the effect size $\eta_p^2 = 0.588$ for testing completed on SVD.

This trend can be attributed to the richer and more varied phonatory patterns present in continuous speech, which capture a broader range of pathophysiological characteristics. In contrast, sustained phonation tends to exhibit more stationary acoustic and bioimpedance patterns, limiting the diversity of features available for classification. Based on these findings, continuous speech was prioritised for all multimodal system developments and evaluations.

10.1.2 Comparison of Feature Representations Performance

An extensive comparison of feature extraction techniques was conducted to identify the most effective representations for each modality. As shown in chapter 8 and 9, as well as the above subsection (*10.1.1 Sustained Phonation versus Continuous Speech*), the highest performance metrics were achieved for the ERB-spectrum-based feature extraction methods – the Gammatone spectrograms and the Gammatone Cepstral Coefficients (GTCCs).

For the audio modality, GTCCs consistently outperformed traditional spectrograms (STFT-derived), Mel-spectrograms, raw waveform approaches, as well as the MFCCs that are widely used for speech-related signal classification (Godino-Llorente and Gomez-Vilda, 2004; Arias-Londoño *et al.*, 2010; Markaki and Stylianou, 2011; Borsky *et al.*, 2017; Wang *et al.*, 2022). For the EGG modality, the Gammatone spectrograms provided the highest performance metrics as compared to other feature extraction methods. For instance: in binary classification, the GTCC-based audio best-performing model ("small" 1D-CNN) achieved an average accuracy over 10-fold cross-validation of 89.17% ± 7.30 compared to 72.02% ± 9.10 achieved with MFCCs. The next-best-performing feature representation for

the audio pathology detection models was speech WAV files (fed into the “big” 1D-CNN), delivering 85.42% ± 5.05, and Gammatone spectrograms (fed into the “small” 1D-CNN), achieving 81.30% ± 2.68 on speech signals from SVD.

In multi-class classification, the EGG-derived Gammatone spectrograms (fed into the “small” 1D-CNN) achieved an average accuracy of 74.21% ± 5.41 calculated over 10-fold cross-validation. The next-best-performing feature representations included raw speech WAV files (73.73% ± 5.01 on the “big” 1D-CNN) and GTCCs (72.90% ± 5.00 on the “small” 1D-CNN). The results lead to the conclusion that biologically inspired ERB-based feature representations outperform other methods in classification of bio-medical signals.

10.1.3 Unimodal versus Multimodal Performance

The designed unimodal systems fed with speech-derived ERB-spectrum-based feature representations demonstrated strong classification capability for both audio and bioimpedance signals. For laryngeal pathology detection (binary classification) performed on the custom dataset (OURs), the best audio-based unimodal system – GTCC features in the “small” 1D-CNN – achieved the average accuracy of 89.17% ± 7.30, while the best EGG-speech-based system – raw WAVs in the 2D-CNN – achieved 86.03% ± 4.94. Similarly, for multi-class classification, the best audio-based system – GTCCs in the “small” 1D-CNN – reached 79.00% ± 7.51, and the EGG system using Gammatone spectrograms attained 74.21% ± 5.41. This confirmed the utility of both modalities in isolation and highlighted audio as more consistent for multi-class tasks – likely due to the richer phonatory variation.

Nevertheless, the multimodal systems showed notable performance improvements, particularly when leveraging stacked generalisation-based late fusion techniques. The early fusion based on audio-derived GTCC and EGG-based Gammatone spectrograms concatenation produced slight gains over unimodal baselines. Hybrid fusion, combining

separate networks for each modality and merging their learned features, improved performance further in both binary and multi-class settings. The late fusion strategy using stacked generalisation to combine final predictions from dedicated unimodal classifiers consistently outperformed all other models. In binary classification, both designed and tested late fusion models achieved the average accuracy calculated over 10-fold cross-validation of >94%, with precision, sensitivity, specificity and F1 scores exceeding those of the best unimodal systems by a wide margin (all reaching >92%). The best performing model for the laryngeal pathology detection was the multimodal late fusion model combining audio-derived GTCCs with EGG-derived Gammatone spectrograms; over 10-fold cross-validation the model achieved: average accuracy of 94.92% ± 2.82, precision of 96.67% ± 2.90, sensitivity of 93.07% ± 3.53, specificity of 96.77% ± 2.84, F1 of 94.81% ± 2.90.

In multi-class laryngeal pathology classification, the stacked generalisation late fusion similarly surpassed other models, maintaining high per-class accuracy, particularly for the class of cancerous and precancerous lesions, which posed the greatest challenge in unimodal setups. ANOVA confirmed a highly significant main effect of the model type on laryngeal pathology classification, specifically on the accuracy of detection of cancerous and precancerous lesions ($F(6,63) = 8.44$, $p < 0.000001$, $\eta_p^2 = 0.446$ for the custom dataset; $F(6,63) = 7.62$, $p < 0.000004$, $\eta_p^2 = 0.420$ for SVD). The best performance for the multi-class laryngeal pathology classification capable of accurately identifying cancerous and precancerous lesions was achieved for the late fusion multimodal model fusing audio-derived GTCCs with EGG-derived Gammatone spectrograms – over 10-fold cross-validation the model reached the average accuracy of 88.00% ± 2.29, as well as the following for the class of cancerous and precancerous lesions: accuracy of 89.23% ± 1.95, precision of 82.32% ± 2.78, sensitivity of 83.27% ± 3.43, specificity of 91.90% ± 1.75, F1 of 82.76% ± 2.60. Tukey's HSD further confirmed that the late fusion multimodal model fusing

audio-derived GTCCs with EGG-derived Gammatone spectrograms significantly outperforms the unimodal baselines in classifying the laryngeal pathologies while detecting the malignancies (mean difference -0.109, 95% CI $= [-0.164, -0.053]$, $p < 0.000003$ while comparing EGG baseline with described late fusion model; mean difference -0.076, 95% CI $= [-0.132, -0.020]$, $p < 0.002$ while comparing it with audio baseline).

Performance was evaluated on both a custom dataset and the public Saarbruecken Voice Database (SVD). While results from OURs remained consistently stronger, especially in multimodal fusion, the SVD experiments confirmed the generalisability trends – particularly the stacking late multimodal fusion system combining audio-derived GTCCs and EGG-derived Gammatone spectrograms outperforming all other investigated models. Thus, notably, the best-performing fusion configurations on the custom dataset also maintained their relative performance on SVD, supporting their robustness across speakers and recording conditions.

10.1.4 Generalisability – Custom Dataset versus SVD Results

Performance evaluation on both the custom dataset (OURs) and the public Saarbruecken Voice Database (SVD) confirmed generalisability trends. Although the custom dataset consistently yielded higher classification scores, the best-performing multimodal fusion models also demonstrated the most accurate and robust performance on SVD.

Overall, the results obtained from the SVD testing confirmed the conclusions of:

- ERB-spectrum-based feature representations (Gammatone spectrograms and GTCCs) outperforming other feature extraction methods,
- Continuous speech outperforming sustained phonation in detection and multi-class classification of laryngeal pathologies,

- Late fusion multimodal approach to classification delivers better classification performance to that obtained by the unimodal approaches.

The study also identified and accounted for limitations of the SVD dataset, including repetitions of the same participants with different ID numbers and limited representation of certain pathology types. These limitations reinforced the value of the custom dataset built for the purposes of this research and designed to better reflect clinical diagnostic needs.

10.2. REVIEW OF RESEARCH OBJECTIVES

The central hypothesis of this research was that a multimodal deep learning model integrating audio and laryngeal bioimpedance signals, utilising Equivalent Rectangular Bandwidth spectrum-based feature extraction and continuous speech instead of sustained phonation, enables superior detection and classification of laryngeal pathologies – particularly cancerous and precancerous lesions – compared to single-modality approaches and systems utilising other feature extraction methods and sustained phonation.

To prove the initial hypothesis, the following were completed:

1. The custom dataset of audio and simultaneously recorded laryngeal bioimpedance was compiled, including recordings of healthy participants, as well as the subjects suffering from various laryngeal pathologies, focusing majorly on cancerous and precancerous lesions.
2. The collected data was subjected to the exploratory data analysis using various statistical and time-frequency parameters, and PCA. The clustering tendencies were assessed using Hopkins statistics and Euclidean distances.
3. Various feature representations were derived from the initial audio and EGG recordings to assess which feature extraction method delivers the best classification performance.

We prove that the ERB-based methods, including Gammatone spectrograms and GTCCs, deliver the highest values of classification performance metrics.

4. Several classification models were developed – including EL, CNN, and RNN architectures – and tested on both audio and laryngeal bioimpedance signals to determine the best-performing classification architectures for the detection and classification of the laryngeal pathologies based on audio and bioimpedance data modalities. Generally, the CNNs outperformed the remaining architectures. The best performing unimodal systems became the building blocks for the subsequently developed multimodal classification system.
5. Both types of phonation recordings – the sustained vowel phonation and the continuous speech – were assessed as the input for the unimodal laryngeal pathology detection and classification systems to conclude which phonation type is more suitable for the accurate and robust detection and classification of laryngeal pathologies. Continuous speech signals outperformed sustained phonation.
6. The several multimodal classification systems were developed and examined, each relying on a different multimodal data fusion strategy, to determine which of the early, hybrid, and late fusion approached performs best for the detection and classification of the laryngeal pathologies. The late fusion approach with the application of the meta-classifier at the decision stage outperformed the remaining systems.
7. The best-performing unimodal and all multimodal classification approaches were compared to determine which strategy delivers the best results for the detection and classification of the laryngeal pathologies. The multimodal classification approach outperformed the unimodal strategy, with late fusion delivering the highest values of classification performance parameters.

Thus, the results obtained in this study strongly support the original hypothesis. The multimodal systems consistently outperformed the unimodal counterparts across both binary and multi-class classification tasks. In particular, the late fusion models achieved the highest classification performance metrics of the accuracy, precision, sensitivity, specificity and F1 scores, validating the advantage of integrating modality-specific networks at the decision level using the ECOC meta-classifier.

It can be stated that a multimodal deep learning model integrating audio and laryngeal bioimpedance signals, utilising Equivalent Rectangular Bandwidth spectrum-based feature extraction and continuous speech instead of sustained phonation, enables superior detection and classification of laryngeal pathologies – particularly cancerous and precancerous lesions – compared to single-modality approaches and systems utilising other feature extraction methods and sustained phonation.

10.3. CONTRIBUTIONS

Through the completion of the objectives outlined in the previous subsection (*10.2 Restatement of Research Hypothesis*), this research has made several novel contributions to the fields of laryngeal pathology classification based on digital audio and laryngeal bioimpedance (electroglottographic – EGG) signals, as well as the multimodal deep learning systems for medical diagnostics. The contributions are summarised as follows:

1. Development of a Standalone Multimodal Dataset for Laryngeal Pathology Classification. A unique dataset was developed comprising of simultaneous recordings of audio and laryngeal bioimpedance (EGG) signals from both healthy participants and individuals with a range of laryngeal pathologies. The dataset contains various phonation types, including sustained vowel phonation and continuous speech of reading a text paragraph. Special emphasis was placed on

collecting data for precancerous and cancerous lesions, addressing a critical gap in existing datasets. This custom dataset provides a valuable foundation for future research and clinical applications.

2. Comprehensive Data Analysis and Identification of Class Separability Patterns. Extensive exploratory data analysis was conducted on the collected dataset, including statistical analysis, time-frequency feature extraction, and principal component analysis (PCA). This analysis revealed patterns of intra- and inter-class variability, demonstrated the greater stability of EGG signals compared to audio, and confirmed that combining modalities enhances class separability. Furthermore, these findings validated the use of deep learning methods for accurate and robust classification of laryngeal pathologies and identification of malignant conditions.
3. Critical Evaluation of Existing Public Datasets. The research highlighted significant limitations within existing resources, particularly the existing datasets (SVD). Issues such as demographic biases, speaker-dependence, and underrepresentation of certain pathologies were identified (particularly malignant conditions), with broader implications for studies relying on these datasets. This work stresses the importance of careful data curation and bias mitigation in medical AI research.
4. Comparative Analysis of Feature Extraction Methods for Pathology Detection and Classification. A systematic evaluation of feature extraction methods for both audio and EGG signals was performed. The study confirmed that features derived from the Equivalent Rectangular Bandwidth (ERB) spectrum, including GTCCs and gammatone spectrograms, significantly outperformed traditional features such as MFCCs and STFT spectrograms. These results contribute to optimising feature selection for future pathology detection systems.

5. Demonstration of Superiority of Continuous Speech over Sustained Phonation for Detection and Classification of Laryngeal Pathologies. Contrary to prevailing trends in the literature, this research demonstrated that continuous speech recordings provide superior diagnostic information compared to sustained vowel phonation. Continuous speech resulted in higher classification accuracies across both unimodal and multimodal systems, highlighting its greater phonatory richness and clinical relevance.
6. Development of Robust Multimodal Laryngeal Pathology Detection and Classification Systems. Accurate and unbiased unimodal classification systems were developed for both audio and EGG modalities, achieving high performance in binary pathology detection and multi-class classification. Building upon these, a series of multimodal systems were created, culminating in a late fusion model that significantly outperformed all unimodal counterparts, thereby validating the hypothesis that multimodal learning enhances diagnostic performance.
7. Comparative Analysis of Multimodal Data Fusion Strategies for Bio-medical Signal Classification. A comprehensive investigation into multimodal fusion approaches – early, hybrid (intermediate), and late fusion – was undertaken. The results demonstrated that late fusion strategies, involving stacked generalisation of modality-specific classifiers, provided the most robust and accurate pathology detection and classification. This work advances the understanding of how best to integrate complementary physiological signals for medical AI applications.
8. Contribution to the Knowledge Base on Medical Electroglottography. Through a dedicated review (Tomaszewska and Georgakis, 2023), this research systematised existing knowledge on the use of electroglottography in medical diagnostics. This

work provides a valuable resource for researchers and clinicians working with laryngeal bioimpedance signals.

10.4. FUTURE WORK DIRECTIONS

While this research has made significant advancements in multimodal laryngeal pathology detection and classification, it opens several avenues for future investigation and system enhancement. The following directions are proposed:

10.4.1 Expansion and Diversification of the Dataset and Classes of Pathologies

Although the custom dataset developed in this study addressed critical gaps, further expansion in terms of participant diversity, pathology types, and linguistic variability would strengthen the generalisability of the classification models. Given AI models benefit from large databases, the number of participants in cancerous and precancerous group should also be expanded. Increasing representation for rarer pathologies and recording speech samples in multiple languages could enhance the system's clinical applicability.

Furthermore, future work will aim at separating the cancerous and precancerous lesions, which was not feasible in this study due to data limitations and the requirement for further invasive diagnostic confirmation. We hope that the enlargement of the dataset and close collaboration with medical experts will enable expansion of the developed pre-screening method into a clinically viable diagnostic tool.

10.4.2 Refinement of Multimodal Feature Engineering

Future work could explore more advanced feature extraction and fusion techniques, including end-to-end learned representations that combine audio and EGG signals at raw or minimally processed levels. Additionally, attention-based fusion mechanisms or graph-

based models could be investigated to dynamically weight modality contributions during classification.

An important direction for future research lies in exploring more advanced representation learning methods, including autoencoders and transfer learning from large pre-trained models. Applying these frameworks to laryngeal audio and bioimpedance signals could potentially enhance diagnostic accuracy and robustness. Nonetheless, many of these approaches function as “black boxes” with limited interpretability, which poses challenges for clinical adoption where biological insight and explainability are crucial. A balanced approach that combines the power of representation learning with transparent, physiologically meaningful features would therefore represent a promising next step.

10.4.3 Real-Time and Embedded System Development

Adapting the developed models for real-time or low-power embedded devices would expand their use beyond clinical settings to mobile health applications. Research into lightweight model architectures and on-device feature extraction could enable the deployment of pathology detection tools on smartphones or portable diagnostic equipment, as well as pre-diagnostic methods for patients anticipating a clinical assessment with a medical professional.

10.4.4 Exploration of Explainability and Interpretability

Improving the interpretability of deep learning-based pathology classifiers remains a critical goal for clinical adoption. Future studies could apply explainable AI techniques to multimodal models to highlight signal regions or features most influential to diagnostic decisions, facilitating model validation.

10.4.5 Investigation of Longitudinal Monitoring

A promising future direction involves applying multimodal models for longitudinal monitoring of patients undergoing treatment for laryngeal disorders. Tracking voice and laryngeal bioimpedance signal changes over time could enable early detection of disease progression or response to therapy, providing valuable support for clinical decision-making.

10.4.6 Clinical Validation Studies

The final and essential step involves conducting clinical validation studies in collaboration with healthcare institutions. Prospective testing with real-world patient populations would provide robust evidence of diagnostic accuracy, usability, and system impact on clinical workflows, paving the way for potential regulatory approval and deployment.

10.5. CLOSING REMARKS

In conclusion, this thesis has demonstrated that the integration of audio recordings and simultaneously gathered laryngeal bioimpedance measurements of speech, supported by carefully designed multimodal deep learning frameworks, offers a significant advancement in the non-invasive detection and classification of laryngeal pathologies. Through rigorous analysis, methodical system development, and critical reflection on the limitations and future possibilities, this work contributes both practical solutions and theoretical insights to the field.

11. REFERENCES

Abberton, E. and Fourcin, A.J. (1972) 'Laryngographic analysis and intonation'. *British Journal of Disorders of Communication*, 7(1), pp.24-29.

Abdi, H. and Williams, L. J. (2010) 'Principal Component Analysis'. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp.433-459.

Abdoli, S., Cardinal, P. and Koerich, A.L. (2019) 'End-to-end environmental sound classification using a 1D convolutional neural network'. *Expert Systems with Applications*, 136, pp.252-263.

Adem, K., Kiliçarslan, S. and Cömert, O. (2019) 'Classification and diagnosis of cervical cancer with softmax classification with stacked autoencoder'. *Expert Systems with Applications*, 115, pp.557-564.

Ahmed, N., Natarajan, T. and Rao, K.R. (1974) 'Discrete cosine transform'. *IEEE Transactions on Computers*, 100(1), pp.90-93.

Al Rahhal, M.M., Bazi, Y., AlHichri, H., Alajlan, N., Melgani, F. and Yager, R.R. (2016) 'Deep learning approach for active classification of electrocardiogram signals'. *Information Sciences*, 345, pp.340-354.

Al-Nasheri, A., Muhammad, G., Alsulaiman, M. and Ali, Z. (2017) 'Investigation of voice pathology detection and classification on different frequency regions using correlation functions'. *Journal of Voice*, 31(1), pp.3-15.

Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T.A., Farahat, M., Malki, K.H. and Bencherif, M.A. (2017) 'An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification'. *Journal of Voice*, 31(1), pp.113-e9.

Alam, M.Z., Simonetti, A., Brillantino, R., Tayler, N., Grainge, C., Siribaddana, P., Nouraei, S.A., Batchelor, J., Rahman, M.S., Mancuzo, E.V. and Holloway, J.W. (2022) 'Predicting pulmonary function from the analysis of voice: a machine learning approach'. *Frontiers in Digital Health*, 4, p.750226.

Altman, K.W. (2007) 'Vocal fold masses'. *Otolaryngologic Clinics of North America*, 40(5), pp.1091-1108.

American Cancer Society (2023) *Cancer Facts & Figures 2023*. Available at: http://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2023-cancer-facts-figures.html (Accessed: 19 October 2023).

Arias-Londoño, J.D., Godino-Llorente, J.I., Sáenz-Lechón, N., Osma-Ruiz, V. and Castellanos-Domínguez, G. (2010) 'Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients'. *IEEE Transactions on Biomedical Engineering*, 58(2), pp.370-379.

Arifianto, D. and Setijono, H. (2002, October) 'Readability improvement for time frequency analysis of pathological speech'. *In Asia-Pacific Conference on Circuits and Systems*, 1, pp. 191-194. IEEE.

Auger, F., Flandrin, P., Gonçalvès, P. and Lemoine, O. (1996) 'Time-frequency toolbox'. *CNRS France-Rice University*, 46.

Aykanat, M., Kılıç, Ö., Kurt, B. and Saryal, S. (2017) 'Classification of lung sounds using convolutional neural networks'. *EURASIP Journal on Image and Video Processing*, 2017(1), pp.1-9.

Baken, R. J. (1992). 'Electroglottography'. *Journal of Voice*, 6(2):98–110.

Barry, W. J. and Putzer, M. (2007) 'Saarbrucken Voice Database. Institute of Phonetics, University of Saarland'. Institute of Phonetics University of Saarland.

Beranek, L. L. (1949) *Acoustic Measurements*. London, United Kingdom: John Wiley.

Berkovsky, S., Cantador, I. and Tikk, D. (2018) *Collaborative Recommendations: Algorithms, Practical Challenges and Applications*. London, UK: World Scientific Publishing Company.

Bishop, P.J., (1980) 'Evolution of the stethoscope'. *Journal of the Royal Society of Medicine*, 73(6), pp.448-456.

Blagouchine, I.V. and Moreau, E. (2011) A'nalytic method for the computation of the total harmonic distortion by the Cauchy method of residues'. *IEEE Transactions on Communications*, 59(9), pp.2478-2491.

Boersma, P. (1993) ‘Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound’. *In Proceedings of The Institute Of Phonetic Sciences,* 17(193), pp. 97-110.

Bohlender, J. (2013) ‘Diagnostic and Therapeutic Pitfalls in Benign Vocal Fold Diseases’. *GMS Current Topics in Otorhinolaryngology, Head and Neck Surgery*, 12, pp. Doc01.

Bonet-Sola, D. and Alsina-Pages, R.M. (2021) ‘A comparative survey of feature extraction and machine learning methods in diverse acoustic environments’. *Sensors*, 21(4), pp.1274.

Borsky, M., Mehta, D.D., Van Stan, J.H. and Gudnason, J. (2017) ‘Modal and nonmodal voice quality classification using acoustic and electroglottographic features’. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), pp.2281-2291.

Bounareli, S., Tzelepis, C., Argyriou, V., Patras, I. and Tzimiropoulos, G. (2025, May) ‘DiffusionAct: Controllable Diffusion Autoencoder for One-shot Face Reenactment’. *In 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG),* pp. 1-11. IEEE.

Brownlee, J. (2017) *Long short-term memory networks with python: develop sequence prediction models with deep learning*. Machine Learning Mastery.

Brownlee, J. (2021) *Ensemble learning algorithms with Python: Make better predictions with bagging, boosting, and stacking*. Machine Learning Mastery.

Carding, P.N., Horsley, I.A. and Docherty, G.J. (1999) ‘A study of the effectiveness of voice therapy in the treatment of 45 patients with nonorganic dysphonia’. *Journal of Voice*, 13(1), pp.72-104.

Chen, L. and Chen, J. (2022) ‘Deep neural network for automatic classification of pathological voice signals’. *Journal of Voice*, 36(2), pp. 288-e15.

Childers, D.G. (1987) ‘Vocal fold vibrations: an EGG model’. *Laryngeal Function in Phonation and Respiration*, pp.181-202.

Childers, D.G., Hicks, D.M., Moore, G.P. and Alsaka, Y.A. (1986) ‘A model for vocal fold vibratory motion, contact area, and the electroglottogram’. *The Journal of the Acoustical Society of America*, 80(5), pp.1309-1320.

Childers, D.G. and Krishnamurthy, A.K. (1985) ‘A critical review of electroglottography’. *Critical reviews in biomedical engineering*, 12(2), pp.131-161.

Childers, D.G. and Larar, J.N. (1984) ‘Electroglottography for laryngeal function assessment and speech analysis’. *IEEE Transactions on Biomedical Engineering*, (12), pp.807-817.

Colton, R.H. and Conture, E.G. (1990) ‘Problems and pitfalls of electroglottography’. *Journal of Voice,* 4(1), pp.10-24.

Courey, M.S., Scott, M.A., Shohet, J.A. and Ossoff, R.H. (1996) ‘Immunohistochemical Characterization of Benign Laryngeal Lesions’. *Annals of Otology, Rhinology & Laryngology*, 105(7), pp.525-531.

Darling, A.M. (1991) ‘Properties and implementation of the gammatone filter: a tutorial’. *Speech Hearing and Language, Work in Progress, University College London, Department of Phonetics and Linguistics*, pp.43-61.

Darvish, M. and Kist, A.M. (2024) ‘A Generative Method for a Laryngeal Biosignal’. *Journal of Voice*, February 2024.

Davis, G., Davis, G. D., Jones, R. (1989) *The Sound Reinforcement Handbook*. United States: Hal Leonard.

De Boer, E. and Kuyper, P. (1968) ‘Triggered correlation’. *IEEE Transactions on Biomedical Engineering*, (3), pp.169-179.

Deshpande, P.S. and Manikandan, M.S. (2017) ‘Effective glottal instant detection and electroglottographic parameter extraction for automated voice pathology assessment’. *IEEE Journal of Biomedical and Health Informatics*, 22(2), pp.398-408.

Dewan, K., Chhetri, D.K. and Hoffman, H. (2022) ‘Reinke's edema management and voice outcomes’. *Laryngoscope – Investigative Otolaryngology*, 7(4), pp.1042-1050.

Dietterich, T.G. and Bakiri, G. (1991, July) ‘Error-correcting output codes: a general method for improving multiclass inductive learning programs’. *In Proceedings of the 9th National Conference on Artificial intelligence-Volume 2*, pp. 572-577.

Donati, E. (2022) *An efficient phonation-driven control system using laryngeal bioimpedance and machine learning* (Doctoral dissertation, University of West London).

DuBois, R.N. (2021) 'The Urgent Need for Expanded Cancer Screening'. *Cancer Prevention Research*, 14(12), pp.1053-1054.
Dworkin, J.P. (2008) 'Laryngitis: types, causes, and treatments'. *Otolaryngologic Clinics of North America*, 41(2), pp.419-436.
Ellis, D. P. W. (2009) 'Gammatone-like spectrograms', web resource. Available at: http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/ [Accessed: 10 November 2024].
Fernandes, J., Teixeira, F., Guedes, V., Junior, A. and Teixeira, J.P. (2018) 'Harmonic to noise ratio measurement-selection of window and length'. *Procedia computer science*, 138, pp.280-285.
Ferrán, S., Garaycochea, O., Terrasa, D., Díaz Zufiaurre, N., Alcalde, J. and Fernández, S. (2024) 'Biography of muscle tension dysphonia: a scoping review'. *Applied Sciences*, 14(5), p.2030.
Ferreira, A.J. and Figueiredo, M.A. (2012) 'Boosting algorithms: A review of methods, theory, and applications'. *Ensemble machine learning: Methods and applications*, pp.35-85.
Fletcher, H. (1940) 'Auditory patterns'. *Reviews of Modern Physics*, 12(1), pp.47.
Fletcher, H. and Munson, W.A. (1933) 'Loudness, its definition, measurement and calculation'. *Bell System Technical Journal*, 12(4), pp.377-430.
Fourcin, A.J. (1974) 'Laryngographic Examination of Vocal Fold Vibration'. *Ventilatory and Phonatory Control Systems*, pp.315-333.
Fourcin A.J., Abberton, E. (1971) 'First applications of a new laryngograph'. *Medical and Biological Illustration*, (21):172-82.
Gadzicki, K, Khamsehashari, R. and Zetzsche, C. (2020) 'Early vs Late Fusion in Multimodal Convolutional Neural Networks'. *IEEE 23rd International Conference on Information Fusion (FUSION)*, Rustenburg, South Africa, pp. 1-6, DOI: 10.23919/FUSION45008.2020.9190246.
Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M. and Ritter, M. (2017, March) 'Audio set: An ontology and human-labeled dataset for audio events'. *In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 776-780). IEEE.
Geng, L., Liang, Y., Shan, H., Xiao, Z., Wang, W. and Wei, M. (2022) 'Pathological voice detection and classification based on multimodal transmission network'. *Journal of Voice*.
Glasberg, B.R. and Moore, B.C. (1990) 'Derivation of auditory filter shapes from notched-noise data'. *Hearing Research*, 47(1-2), pp.103-138.
Godino-Llorente, J.I. and Gómez-Vilda, P. (2004) 'Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors'. *IEEE Transactions on Biomedical Engineering*, 51(2), pp.380-384.
Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. London, UK: MIT Press.
Graupe, D. (2013) *Principles of Artificial Neural Networks*. 3rd edition. Singapore: World Scientific.
Gregory, N.D., Chandran, S., Lurie, D. and Sataloff, R.T. (2012) 'Voice disorders in the elderly'. *Journal of Voice*, 26(2), pp.254-258.
Groeneveld, R.A. and Meeden, G., (1984) 'Measuring skewness and kurtosis'. *Journal of the Royal Statistical Society Series D: The Statistician*, 33(4), pp.391-399.
Grzywalski, T., Piecuch, M., Szajek, M., Bręborowicz, A., Hafke-Dys, H., Kociński, J., Pastusiak, A. and Belluzzo, R. (2019) 'Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination'. *European Journal of Pediatrics*, 178, pp.883-890.
Hansen, J. H. L. and Patil, S. (2007) 'Speech Under Stress: Analysis, Modeling and Recognition'. *Lecture Notes in Computer Science*, 4343, pp. 108–137.
Harar, P., Alonso-Hernandezy, J.B., Mekyska, J., Galaz, Z., Burget, R. and Smekal, Z. (2017, July) 'Voice pathology detection using deep learning: a preliminary study'. *In 2017 international conference and workshop on bioinspired intelligence (IWOBI)* (pp. 1-4). IEEE.
Hemmerling, D., Skalski, A. and Gajda, J. (2016) 'Voice data mining for laryngeal pathology assessment', *Computers in Biology and Medicine*, 69(1), pp. 270-276, Feb. 2016.
Henrich, N., d'Alessandro, C., Doval, B. and Castellengo, M. (2004) 'On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation'. *The Journal of the Acoustical Society of America,* 115(3), pp.1321-1332.

Henrich, N., d'Alessandro, C., Doval, B. and Castellengo, M. (2005) 'Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency'. *The Journal of the Acoustical Society of America*, 117(3), pp.1417-1430.

Henríquez, P., Alonso, J.B., Ferrer, M.A., Travieso, C.M., Godino-Llorente, J.I. and Díaz-de-María, F. (2009) 'Characterization of healthy and pathological voice through measures based on nonlinear dynamics'. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6), pp.1186-1195.

Herbst, C.T. (2020) 'Electroglottography–an update'. *Journal of Voice*, 34(4), pp.503-526.

Herbst, C.T. and Dunn, J.C. (2019) 'Fundamental frequency estimation of low-quality electroglottographic signals'. *Journal of Voice*, 33(4), pp.401-411.

Ho, T.K. (1995, August) 'Random decision forests'. *In Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1, pp. 278-282. IEEE.

Holdsworth, J., Nimmo-Smith, I., Patterson, R. and Rice, P. (1988) 'Implementing a gammatone filter bank'. *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, 1, pp.1-5.

Hopkins, B. and Skellam, J.G. (1954) 'A new method for determining the type of distribution of plant individuals'. *Annals of Botany*, 18(2), pp.213-227.

Hosokawa, K., Ogawa, M., Hashimoto, M. and Inohara, H. (2014) 'Statistical analysis of the reliability of acoustic and electroglottographic perturbation parameters for the detection of vocal roughness'. *Journal of Voice*, 28(2), pp.263-e9.

Imai, S. (1983, April) 'Cepstral analysis synthesis on the mel frequency scale'. *In ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 8, pp. 93-96). IEEE.

Islam, R., Abdel-Raheem, E. and Tarique, M. (2022, November) 'Deep Learning Based Pathological Voice Detection Algorithm Using Speech and Electroglottographic (EGG) Signals'. *In 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)* (pp. 127-131). IEEE.

Islam, R., Abdel-Raheem, E. and Tarique, M. (2022) 'Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals'. *Computer Methods and Programs in Biomedicine Update*, 2, p.100074.

Jetté, M. (2016) 'Toward an Understanding of the Pathophysiology of Chronic Laryngitis'. *Perspectives of the ASHA Special Interest Groups*, 1(3), pp.14-25.

Jiang, J.J., Tang, S., Dalal, M., Wu, C.H. and Hanson, D.G. (1998) 'Integrated analyzer and classifier of glottographic signals'. *IEEE Transactions on Rehabilitation Engineering*, 6(2), pp.227-234.

Johannesma, P.L.M. (1972) 'The pre-response stimulus ensemble of neurons in the cochlear nucleus'. *In Symposium on Hearing Theory (IPO, Eindhoven, Holland)*, pp. 58-69.

Johns, M.M., (2003) 'Update on the etiology, diagnosis, and treatment of vocal fold nodules, polyps, and cysts'. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 11(6), pp.456-461.

Johnston, J. D. (1988) 'Transform Coding of Audio Signals Using Perceptual Noise Criteria'. *IEEE Journal on Selected Areas in Communications*. 6(2), pp. 314–323.

Jones, T. M., De, M., Foran, B., Harrington, K., & Mortimore, S. (2016) 'Laryngeal cancer: United Kingdom national multidisciplinary guidelines'. *The Journal of Laryngology & Otology*, 130(S2), S75-S82.

Kim, H., Moreau, N., Sikora, T. (2006) *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. West Sussex, England: John Wiley & Sons, Ltd.

Kim, P. (2017) *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*. Seoul, Korea: Apress.

Kitzing, P. (1985) 'Stroboscopy - a pertinent laryngological examination'. *The Journal of Otolaryngology*, 14(3), pp.151-157.

Kumar, D., Satija, U. and Kumar, P. (2023, February) 'Analysis and Classification of Electroglottography Signals for the Detection of Speech Disorders'. In *2023 National Conference on Communications (NCC)* (pp. 1-6). IEEE.

Kuo, H.C., Hsieh, Y.P., Tseng, H.H., Wang, C.T., Fang, S.H. and Tsao, Y. (2023) 'Toward Real-World Voice Disorder Classification'. *IEEE Transactions on Biomedical Engineering*.

Lahat, D., Adali, T. and Jutten, C. (2015) 'Multimodal data fusion: an overview of methods, challenges, and prospects'. *Proceedings of the IEEE*, 103(9), pp.1449-1477.

Lecluse, F.L.E., Brocaar, M.P. and Verschuure, J. (1975) 'The electroglottography and its relation to glottal activity'. *Folia Phoniatrica et Logopaedica*, 27(3), pp.215-224.
LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, DOI: 10.1109/5.726791.
Lee, J. Y. (2021) 'Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database'. *Applied Sciences*, 11(15), p.7149.
Lee, C. and Landgrebe, D.A. (1993) 'Feature extraction based on decision boundaries'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4), pp.388-400.
Lee, W., Seong, J.J., Ozlu, B., Shim, B.S., Marakhimov, A. and Lee, S. (2021) 'Biosignal sensors and deep learning-based speech recognition: A review'. *Sensors,* 21(4), p.1399.
Lee, Y., Kim, G., Wang, S., Jang, J., Cha, W., Choi, H. and Kim, H. (2019) 'Acoustic characteristics in epiglottic cyst'. *Journal of Voice*, 33(4), pp.497-500.
Lerch, A. (2012) *An Introduction to Audio Content Analysis Applications in Signal Processing and Music Informatics*. Piscataway, New Jersey, USA: IEEE Press.
Li, Y. and Wang, D. (2007) 'Separation of singing voice from music accompaniment for monaural recordings'. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), pp.1475-1487.
Liu, G.S., Hodges, J.M., Yu, J., Sung, C.K., Erickson-DiRenzo, E. and Doyle, P.C. (2023) 'End-to-end deep learning classification of vocal pathology using stacked vowels'. *Laryngoscope Investigative Otolaryngology*, 8(5), pp.1312-1318.
Liu, M., Zhang, D., Chen, S. and Xue, H. (2015) 'Joint binary classifier learning for ECOC-based multi-class classification'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11), pp.2335-2341.
Lyberg-Ahlander, V., Rydell, R., Fredlund, P., Magnusson, C. and Wilén, S. (2019) 'Prevalence of voice disorders in the general population, based on the Stockholm public health cohort'. *Journal of Voice*, 33(6), pp.900-905.
Lyon, R.F., Katsiamis, A.G. and Drakakis, E.M. (2010) 'History and future of auditory filter models'. *In Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 3809-3812. IEEE.
Marchal, A. (2009) *From Speech Physiology to Linguistic Phonetics*. London, UK: John Wiley & Sons, Inc.
Markaki, M. and Stylianou, Y. (2011) 'Voice pathology detection and discrimination based on modulation spectral features'. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), pp.1938-1948.
Markatopoulou, F., Mezaris, V. and Patras, I. (2018) 'Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation'. *IEEE Transactions on Circuits and Systems for Video Technology, 29*(6), pp.1631-1644.
Martins, R. H. G., do Amaral, H. A., Tavares, E. L. M., Martins, M. G., Gonçalves, T. M., & Dias, N. H. (2015) 'Voice disorders: Etiology and diagnosis'. *Journal of Voice*, 30(6), 761.e1–761.e9.
Martins, R. H. G., Tavares, E. L. M., Ranalli, P. F., Branco, A. and Pessin, A.B.B. (2014) 'Psychogenic dysphonia: diversity of clinical and vocal manifestations in a case series'. *Brazilian Journal of Otorhinolaryngology*, 80(6), pp.497-502.
Massachusetts Eye and Ear Infirmary (1994) '*Voice disorders database, version. 1.03 (cd-rom)*'. Lincoln Park, NJ: Kay Elemetrics Corporation.
MathWorks Inc. (2024) '*Help Center: Spectral Descriptors'*. Available at: https://uk.mathworks.com/help/audio/ug/spectral-descriptors.html [Accessed: 10 November 2024]
McCulloch, W. and Pitts, W. (1943) 'A Logical Calculus of Ideas Immanent in Nervous Activity"', *Bulletin of Mathematical Biophysics*, 5 (4): 115–133. doi:10.1007/BF02478259.
Miliaresi, I., Pikrakis, A. and Poutos, K. (2022, September) 'A Deep Multimodal Voice Pathology Classifier with Electroglottographic Signal Processing Capabilities'. *In 2022 7th International Conference on Frontiers of Signal Processing (ICFSP)* (pp. 109-113). IEEE.

Misra, H., Ikbal, S. Bourlard, H. and Hermansky, H. (2004) 'Spectral Entropy Based Feature for Robust ASR'. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE.
Molau, S., Pitz, M., Schluter, R. and Ney, H. (2001, May) 'Computing mel-frequency cepstral coefficients on the power spectrum'. *In Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1, pp. 73-76. IEEE.
Moore, B.C. and Glasberg, B.R. (1987) 'Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns'. *Hearing Research*, 28(2-3), pp.209-225.
Moran, R.J., Reilly, R.B., de Chazal, P. and Lacy, P.D. (2006) 'Telephony-based voice pathology assessment using automated speech analysis'. *IEEE Transactions on Biomedical Engineering*, 53(3), pp.468-477.
Moussavi, Z. (2006) *Fundamentals of Respiratory Sounds and Analysis*. Manitoba, Canada: Morgan & Claypool Publishers.
Mohammed, H.M., Omeroglu, A.N. and Oral, E.A. (2023) 'MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection'. *Expert Systems with Applications*, 223, p.119790.
Mohammed, M.A., Abdulkareem, K.H., Mostafa, S.A., Khanapi Abd Ghani, M., Maashi, M.S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H. and Al-Dhief, F.T. (2020) 'Voice pathology detection and classification using convolutional neural network model'. *Applied Sciences*, 10(11), p.3723.
Muhammad, G. and Alhussein, M. (2021) 'Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection'. *IEEE Access*, 9, pp.89198-89209.
Myssiorek, D. (2004) 'Recurrent laryngeal nerve paralysis: anatomy and etiology'. *Otolaryngologic clinics of North America*, 37(1), pp.25-44.
Nacci, A., Macerata, A., Bastiani, L., Paludetti, G., Galli, J., Marchese, M.R., Barillari, M.R., Barillari, U., Laschi, C., Cianchetti, M. and Manti, M. (2020) 'Evaluation of the electroglottographic signal variability in organic and functional dysphonia'. *Journal of Voice*, 36(6), pp.881-e5.
Nacci, A., Romeo, S.O., Cavaliere, M.D., Macerata, A., Bastiani, L., Paludetti, G., Galli, J., Marchese, M.R., Barillari, M.R., Barillari, U. and Berrettini, S. (2019) 'Comparison of electroglottographic variability index in euphonic and pathological voice'. *Acta Otorhinolaryngologica Italica*, 39(6), p.381.
National Health Service (2023) *Long-term effects of COVID-19 (long COVID)*. Available at: https://www.nhs.uk/conditions/covid-19/long-term-effects-of-covid-19-long-covid/ (Accessed: 16 October 2023).
National Health Service (2022) *Your Covid Recovery. Swallowing and changes to your voice.* Available at: https://www.yourcovidrecovery.nhs.uk/i-think-i-have-long-covid/effects-on-your-body/swallowing-and-changes-to-your-voice/ (Accessed: 16 October 2023).
Netter, F. H. (2019) *Atlas of Human Anatomy*. 7th edn. Philadelphia, USA: Elsevier Health Sciences.
On, C.K., Pandiyan, P.M., Yaacob, S. and Saudi, A. (2006) 'Mel-frequency cepstral coefficient analysis in speech recognition'. *In 2006 International Conference on Computing & Informatics*, pp. 1-5. IEEE.
Oppenheim, A.V. and Schafer, R.W. (1989) Discrete-Time Signal Processing. New Jersey, USA: Prentice Hall.
Patterson, J. and Gibson, A. (2017) *Deep Learning: A Practitioner's Approach*. Boston, USA: O'Reilly Media.
Patterson, R.D. and Holdsworth, J. (1996) 'A functional model of neural activity patterns and auditory images'. *Advances in Speech, Hearing and Language Processing*, 3(Part B), pp. 547-563.
Patterson R.D. and Moore B.J.C. (1986) 'Auditory filters and excitation patterns as representations of frequency resolution', *In: Frequency Selectivity in Hearing*, B.C.J. Moore (Ed.), Academic Press Patterson.
Patterson, R.D., Nimmo-Smith, I., Holdsworth, J. and Rice, P. (1987) 'An efficient auditory filterbank based on the gammatone function'. *In a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, 2(7).

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J. and Rice, P. (1988) 'Spiral Vos Final Report', *Part A: The Auditory Filter bank (Annex C), Contract Report, Cambridge Electronic Design*. APU rep. 2341.
Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand M. (1992) 'Complex sounds and auditory images'. *In: Auditory physiology and perception, Proc. 9th International Symposium on Hearing*, Eds: Y Cazals, L. Demany, and K. Horner. Pergamon, Oxford, pp.429-446.
Peeters, G. (2004) 'A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project'. *Technical Report; IRCAM*: Paris, France.
Peng, C., Chen, W., Zhu, X., Wan, B. and Wei, D. (2007, October) 'Pathological voice classification based on a single Vowel's acoustic features'. *In 7th IEEE International Conference on Computer and Information Technology (CIT 2007)*, pp.1106-1110. IEEE.
Peterson, W. W. and Weldon, E.J. (1972) *Error-correcting codes*. MIT press: Cambridge, MA, USA.
Pour, A.F., Asgari, M. and Hasanabadi, M.R. (2014) 'Gammatonegram based speaker identification'. *In 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 52-55. IEEE.
Praveen, G.B., Agrawal, A., Sundaram, P. and Sardesai, S. (2018) 'Ischemic stroke lesion segmentation using stacked sparse autoencoder'. *Computers in Biology and Medicine*, 99, pp.38-52.
Przysiezny, P.E. and Przysiezny, L.T.S. (2015) 'Work-related voice disorder'. *Brazilian Journal of Otorhinolaryngology*, 81, pp.202-211.
Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.Y. and Sainath, T. (2019) 'Deep learning for audio signal processing'. *IEEE Journal of Selected Topics in Signal Processing*, 13(2), pp.206-219.
Pützer, M. and Koreman, J. (1997) 'A German database of patterns of pathological vocal fold vibration'. *Phonus*, 3, pp.143-153.
Ramachandram, D. and Taylor, G.W. (2017) 'Deep multimodal learning: A survey on recent advances and trends'. IEEE Signal Processing Magazine, 34(6), pp.96-108.
Rao, A., Huynh, E., Royston, T.J., Kornblith, A. and Roy, S. (2018) 'Acoustic methods for pulmonary diagnosis'. *IEEE Reviews in Biomedical Engineering*, 12, pp.221-239.
Redford, M. A. (2019) *The Handbook of Speech Production*. Oxford, UK: John Wiley & Sons, Inc.
Ritchings, R.T., McGillion, M.A., Conroy, G.V. and Moore, C.J. (1999, October) 'Objective assessment of pathological voice quality'. *In IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)* (Vol. 6, pp. 340-345). IEEE.
Rosa, M.O., Pereira, J.C., Greller, M. and Carvalho, A.C.P.D.L.F.D. (1999, September) 'Signal processing and statistical procedures to identify laryngeal pathologies'. *In ICECS'99. Proceedings of ICECS'99. 6th IEEE International Conference on Electronics, Circuits and Systems*,. 1, pp. 423-426. IEEE.
Rosen, C.A. and Murry, T. (2000) 'Diagnostic laryngeal endoscopy'. *Otolaryngologic Clinics of North America*, 33(4), pp.751-757.
Rossing, T. D. (2007) *Springer Handbook of Acoustics*. Stanford, USA: Springer.
Rothenberg, M. (1992) 'A multichannel electroglottograph'. *Journal of Voice*, *6*(1), pp.36-43.
Rothenberg, M. (1981) 'Some relations between glottal air flow and vocal fold contact area'. *Asha Rep*, 11, pp.88-96.
Roy, N. (2003) 'Functional dysphonia'. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 11(3), 144-148.
Roy, N., Merrill, R.M., Gray, S.D. and Smith, E.M. (2005) 'Voice disorders in the general population: prevalence, risk factors, and occupational impact'. *The Laryngoscope*, 115(11), pp.1988-1995.
Rubin, A.D. and Sataloff, R.T. (2007) 'Vocal fold paresis and paralysis'. *Otolaryngologic Clinics of North America*, 40(5), pp.1109-1131.
Saarbruecken Voice Database: Handbook. Available at: https://stimmdb.coli.uni-saarland.de/help_en.php4 (Accessed: 15 October 2023).
Sarkar, M., Madabhavi, I., Niranjan, N. and Dogra, M. (2015) 'Auscultation of the respiratory system'. *Annals of Thoracic Medicine*, 10(3), p.158.

Sataloff, R. T. (2017) *Vocal Health and Pedagogy: Science, Assessment, and Treatment*. 3rd edn. San Diego, USA: Plural Publishing, Inc.

Scheirer, E., and Slaney, M. (1997) 'Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator'. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE.

Shrivas, A., Deshpande, S., Gidaye, G., Nirmal, J., Ezzine, K., Frikha, M., Desai, K., Shinde, S., Oza, A.D., Burduhos-Nergis, D.D. and Burduhos-Nergis, D.P. (2022) 'Employing energy and statistical features for automatic diagnosis of voice disorders'. *Diagnostics*, 12(11), p.2758.

Slaney, M. (1993) 'An efficient implementation of the Patterson-Holdsworth auditory filter bank'. *Apple Computer, Perception Group, Tech. Rep*, 35(8).

Slaney, M. (1998) 'Auditory Toolbox'. *Interval Research Corporation, Tech. Rep*, 10(1998), pp.1194.

Smith, A.M. and Childers, D.G. (1983) 'Laryngeal evaluation using features from speech and the electroglottograph'. *IEEE Transactions on Biomedical Engineering*, (11), pp.755-759.

Smith, J.O. and Abel, J.S. (1999) 'Bark and ERB bilinear transforms'. *IEEE Transactions on Speech and Audio Processing*, 7(6), pp.697-708.

Soni, H.D., Gandhi, S., Goyal, M. and Shah, U. (2016) 'Study of Clinical Profile of Benign Laryngeal Lesions'. International Journal of Medical Science and Public Health, 5(4), pp.656-660.

Stahlschmidt, S.R., Ulfenborg, B. and Synnergren, J. (2022) 'Multimodal deep learning for biomedical data fusion: a review'. *Briefings in Bioinformatics*, 23(2), p.1-15.

Stevens, S.S., Volkmann, J. and Newman, E.B. (1937) 'A scale for the measurement of the psychological magnitude pitch'. *The Journal of The Acoustical Society of America*, 8(3), pp.185-190.

Tappert, C. C. (2019) 'Who Is the Father of Deep Learning?', *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 343-348, DOI: 10.1109/CSCI49370.2019.00067.

Thomas, M.R. and Naylor, P.A. (2009) 'The SIGMA algorithm: A glottal activity detector for electroglottographic signals'. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8), pp.1557-1566.

Titze, I.R. (1990) 'Interpretation of the electroglottographic signal'. *Journal of Voice*, 4(1), pp.1-9.

Tomaszewska, J.Z., Chousidis, C. and Donati, E. (2022, June) 'Sound-Based Cough Detection System using Convolutional Neural Network'. *In 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6).* IEEE.

Tomaszewska, J.Z., Chousidis, C. and Georgakis, A. (2024, September) 'Comparative Analysis of MFCC and GTCC Performance in Laryngeal Pathology Detection Based on Electroglottographic Signals'. *Institute of Acoustics, 2024, September*, 46(2). DOI: 10.25144/23671

Tomaszewska, J.Z. and Georgakis, A. (2023) 'Systematic Review of Electroglottography in Diagnostics, with Emphasis on Its Implementation in Digital Vocal Tract Pathology Classification Systems'. *Journal of Voice*, December 2023.

Valero, X. and Alias, F. (2012) 'Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification'. *IEEE Transactions on Multimedia*, 14(6), pp.1684-1689.

Van Michel, C.L., Pfister, K.A. and Luchsinger, R. (1970) 'Electroglottographie et cinématographie laryngée ultra-rapide: Comparaison des résultats'. *Folia Phoniatrica et Logopaedica*, 22(2), pp.81-91.

Vanderpump, M.P. (2011) 'The epidemiology of thyroid disease'. *British Medical Bulletin*, 99(1).

Villa-Cañas, T., Arias-Londoño, J.D., Vargas-Bonilla, J.F. and Orozco-Arroyave, J.R. (2015, September) 'Time-frequency approach in continuous speech for detection of Parkinson's disease'. *In 2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*, pp. 1-6. IEEE.

Wang, H., Subramanian, V. and Syeda-Mahmood, T. (2021, April) 'Modeling uncertainty in multi-modal fusion for lung cancer survival analysis'. *In 2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1169-1172. IEEE.

Wang, S.S., Wang, C.T., Lai, C.C., Tsao, Y. and Fang, S.H. (2022) 'Continuous speech for improved learning pathological voice disorders'. *IEEE Open Journal of Engineering in Medicine and Biology*, 3, pp.25-33.

Wang, Z., Müller, M., Caffier, F. and Caffier, P.P. (2023) 'Harnessing Machine Learning in Vocal Arts Medicine: A Random Forest Application for "Fach" Classification in Opera'. *Diagnostics*, 13(18), p.2870.

Wikimedia Commons - Illu conducting passages.svg (2010) Available at: https://commons.wikimedia.org/wiki/File:Illu_conducting_passages.svg [Accessed: February 2022]

Wilkins, R.L., Dexter, J.R., Murphy, R.L. and DelBono, E.A. (1990) 'Lung sound nomenclature survey'. *Chest*, 98(4), pp.886-889.

Winslow, T. (2012) *Larynx Anatomy 2012*. Available at: https://www.teresewinslow.com/#/head/ [Accessed: February 2022].

Wolpert, D.H. (1992) 'Stacked generalization'. *Neural Networks*, 5(2), pp.241-259.

Xu, X., Zhao, X., Xu, W., Cao, J. and Zhang, X. (2023, December) 'Laryngeal image dataset automatic annotation and classification of laryngeal disease'. *In 2023 5th International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, pp. 723-728. IEEE.

Yan, F., Yan, B. and Pei, M. (2023, October) 'Dual transformer encoder model for medical image classification'. *In 2023 IEEE International Conference on Image Processing (ICIP)*, pp. 690-694. IEEE.

Yao, Y., Powell, M., White, J., Feng, J., Fu, Q., Zhang, P. and Schmidt, D.C. (2023) 'A multi-stage transfer learning strategy for diagnosing a class of rare laryngeal movement disorders'. *Computers in Biology and Medicine*, 166, p. 107534.

Yin, L., Liu, Y., Pei, M., Li, J., Wu, M. and Jia, Y., 2021. 'Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism'. *Pattern Recognition Letters*,150, pp.207-213.

Zhou, C., Wu, Y., Fan, Z., Zhang, X., Wu, D. and Tao, Z. (2022) 'Gammatone spectral latitude features extraction for pathological voice detection and classification,' *Applied Acoustics*, 185, p. 108417, Jan. 2022.

12. APPENDICES

12.1. DATA ANALYSIS PARAMETERS OF AUDIO AND EGG SIGNALS – TABLE

Table 12.1: Data analysis parameters calculated for audio and electroglottographic signals of the custom dataset obtained during the exploratory data analysis.

DATA	Params	Mean	Std	Skewness	Kurtosis	Harmonic-to-Noise	Autocorr-Mean	Autocorr-Std	modSpec-Mean
AUDIO Malignant	Mean	-3.2751E-19	1	0.131202541	2.78126641	21.22672572	0.163699924	0.23032169	5.227152874
AUDIO Malignant	Std	1.0877E-17	9.8126E-15	0.4563096	1.2642216	5.817804104	0.152520137	0.08933531	2.088296735
AUDIO Malignant	Min	-9.1398E-17	1	-1.13593408	1.47151844	7.673561462	-0.138884516	0.03819473	1.782813017
AUDIO Malignant	Max	1.2712E-16	1	1.709111627	14.8096916	33.99341626	0.496627446	0.56409979	12.66445036
AUDIO Neuromuscular	Mean	-4.4014E-19	1	0.030946129	2.47571244	22.20518095	0.21147519	0.25144513	4.9010399
AUDIO Neuromuscular	Std	2.0906E-17	1.0432E-14	0.385557223	1.04393498	5.651480435	0.161058641	0.09168652	2.308250201
AUDIO Neuromuscular	Min	-3.6061E-16	1	-1.16162059	1.44986751	4.561233703	-0.079317071	0.01847056	1.90276648
AUDIO Neuromuscular	Max	1.4236E-16	1	1.228680974	9.43700317	33.71157166	0.499446249	0.51982204	16.20818635
AUDIO Healthy	Mean	4.263E-19	1	-0.07512896	2.42274064	23.72874228	0.229280242	0.2535475	3.899410915
AUDIO Healthy	Std	1.1372E-17	1.2333E-14	0.474778288	0.69969786	5.008565632	0.143022471	0.06793796	1.017774526
AUDIO Healthy	Min	-7.8144E-17	1	-1.72490793	1.39084652	11.27387901	-0.061505635	0.07475717	2.018438804
AUDIO Healthy	Max	2.0296E-16	1	1.233617926	5.58089434	40	0.494478692	0.5349771	7.174211772
EGG Malignant	Mean	4.7096E-19	1	0.335264373	2.12042478	23.47032017	0.319422056	0.27474345	4.606026894
EGG Malignant	Std	1.251E-17	1.5311E-14	0.438461946	1.683239	5.931178552	0.168370242	0.06585073	2.021679005
EGG Malignant	Min	-1.1961E-16	1	-1.27905124	1.25832426	3.601430872	-0.168353652	0.02313186	2.520027449
EGG Malignant	Max	1.2956E-16	1	4.070930766	29.4569466	32.1074809	0.498429593	0.49306474	19.01591142
EGG Neuromuscular	Mean	-1.2321E-18	1	0.319749661	2.06782039	21.50989843	0.275817656	0.27126633	5.710396062
EGG Neuromuscular	Std	1.5616E-17	1.3907E-14	0.384608095	0.72946035	7.423963378	0.180741079	0.08153233	3.421379621
EGG Neuromuscular	Min	-2.691E-16	1	-1.00951563	1.22149894	3.603052772	-0.147981546	0.01188141	2.089807497
EGG Neuromuscular	Max	2.6726E-17	1	1.629449516	9.22726533	32.6392524	0.49365493	0.50485138	19.83494342
EGG Healthy	Mean	1.5596E-19	1	0.353169858	1.58856252	26.79740592	0.443314599	0.28028567	3.643274851
EGG Healthy	Std	6.5083E-18	1.4605E-14	0.291663133	0.30304949	3.551457293	0.059097948	0.01877357	0.550240901
EGG Healthy	Min	-2.6183E-17	1	-0.41074557	1.2080302	18.13605472	0.071751341	0.18883622	2.526432369
EGG Healthy	Max	3.3448E-17	1	1.681081589	4.46111043	40	0.495524318	0.40638819	6.075585043

modSpec-Std	SNR	Total-Harmonic-Dist	HarmonicRatio-Mean	HarmonicRatio-Std	sCentroid	centStd	sEntropy	entrStd	sRolloffPoint	rollStd	sSpread	spreadStd
26.2205646	7.464619414	-11.25397587	0.718860206	0.029638572	644.380724	114.046025	0.32249171	0.03567472	2094.50457	593.287826	1079.22653	217.879822
0.79552942	9.273527204	9.738864852	0.146316921	0.039756071	434.744382	186.960968	0.11403388	0.0231298	2106.47599	736.809602	601.020041	145.10561
17.3269613	-10.0671286	-38.85286666	0.152500213	0.00062516	192.554508	3.61105676	0.05196983	0.00498978	272.290827	0	229.594981	33.6963372
27.3056943	35.01835944	5.127002438	0.948122939	0.316152246	3384.15705	2222.68739	0.60669502	0.17327461	10160.8934	4433.25619	3422.15107	1489.63397
26.2817133	9.568695387	-10.71543951	0.752219214	0.024766047	659.220962	92.8638433	0.31544224	0.03057305	1829.82611	439.064488	971.062311	197.883245
0.79455294	9.05683262	8.623699799	0.167280669	0.037769148	677.345089	106.88209	0.11844544	0.02001527	1902.212	539.259169	533.564839	100.868488
21.3630557	-13.7084539	-36.61584983	0.187197563	0.000287676	171.769705	3.92316481	0.04884797	0.00293995	225.056704	0	245.447014	18.9789019
27.4336177	31.12299781	4.538889493	0.937224535	0.287838347	6379.1926	907.207413	0.74498613	0.17386697	11305.6263	3282.39019	3615.64154	880.639845
26.5638052	8.57147469	-10.89378583	0.847588137	0.007278143	627.727656	46.5770034	0.28014274	0.02320901	1305.08455	172.67088	658.281386	150.758855
0.23641216	9.179348871	8.822995636	0.114759438	0.011048637	324.791936	34.1308893	0.10379193	0.01358355	866.036479	199.828431	263.492986	81.5798583
25.4506668	-5.29420801	-34.58668435	0.451629196	0.000115135	167.778551	2.12442589	0.01634946	0.00219433	261.176915	0	214.628288	13.862637
27.4357628	32.510444	4.09589619	0.971631165	0.112098532	2212.49705	544.327916	0.55353118	0.09706026	5315.22807	1414.77661	1964.23586	427.667509
26.3422915	15.33906017	-10.01557499	0.71388245	0.030787807	410.869706	92.2313521	0.24505226	0.03359745	1240.20599	415.690876	958.958135	211.276727
0.80310614	7.162136124	4.752676098	0.155306989	0.047173502	455.593278	191.279991	0.09233413	0.02775239	1985.03462	762.029905	650.67799	174.479793
19.1925344	-12.7639272	-27.98358585	0.117853426	0.00042271	125.171068	0.82426467	0.03963838	0.00154389	172.265625	0	167.118489	4.83696531
27.9466428	28.08804644	-0.042574357	0.947154794	0.300608271	5000.13873	1992.59362	0.8091955	0.25606392	13678.4463	5825.18695	4336.14933	1547.58929
25.9929482	12.73211082	-10.72943029	0.706511024	0.034061212	668.796477	118.466332	0.28322331	0.0341044	2353.90208	490.256704	1285.54856	205.224913
1.42987871	8.515825077	5.861284406	0.204022374	0.051528497	888.108322	191.33797	0.13601235	0.02705888	3371.80866	682.908727	1007.04116	144.408389
17.9834958	-15.67346	-38	0.109922958	0.000417543	128.588794	1.97839261	0.02751902	0.0020944	172.265625	0	308.014641	9.41720194
27.5550615	29.12362199	1.977880449	0.947926507	0.318045052	7608.91915	1722.98287	0.8413375	0.20396712	14328.6101	5383.53898	5454.23306	1317.81547
26.6195766	19.26684165	-8.199131935	0.85205586	0.003418199	348.388685	31.5771537	0.19286957	0.017685	741.886012	109.603347	576.401905	137.355371
0.12951577	2.404623284	2.012155187	0.107841854	0.004221575	118.022893	23.0332198	0.07246364	0.01194653	270.681845	109.013159	175.643099	81.1680807
25.1573225	9.34364848	-16.86320562	0.567070266	0.000191028	139.437851	0.64538782	0.06138858	0.00157612	272.290827	0	153.012279	2.70620132
26.9627452	27.17133503	-1.049512572	0.970701282	0.046437103	689.47548	212.145101	0.3351115	0.06809688	1989.39012	1867.36187	1238.23824	619.377519

sSkewness	skewStd	sKurtosis	kurtStd	sFlatness	flatStd	sCrest	crentStd	sFlux	fluxStd	sSlope	slopeStd	sDecrease	decreaseStd
10.58042483	2.4668213	262.326295	154.648279	0.0086956	0.00623786	122.119249	16.8894203	0.17725643	0.09556739	-4.6067E-06	5.9228E-07	0.116595039	0.04287378
7.627823796	3.16563112	438.343359	393.632257	0.00691317	0.00388038	50.8041823	9.16773982	0.09490093	0.08524919	9.3018E-07	4.6922E-07	0.037453001	0.02629541
0.334871732	0.17280421	2.38795143	0.75786231	0.00043594	0.00033726	41.4423441	1.73960164	0.01578553	0.00867686	-7.1789E-06	6.6709E-08	-0.05584722	0.00378005
51.89111374	21.5784994	3638.5812	3325.9027	0.08135164	0.03535432	243.51133	63.3822658	0.61368851	0.82490384	-2.4284E-06	4.0493E-06	0.26546415	0.18780438
10.70831693	2.43406412	261.674971	146.138096	0.00830831	0.00585816	126.241452	15.6330347	0.16242081	0.08136218	-4.5918E-06	5.4945E-07	0.109828722	0.04122384
7.14349831	2.66980852	406.200342	324.310428	0.00784343	0.00384001	49.3941256	8.81583804	0.09773065	0.06235261	9.1935E-07	4.092E-07	0.037029546	0.03075856
-0.02303708	0.14205249	2.42833648	0.33548477	0.00030344	0.0001678	22.6048325	1.1974076	0.01054442	0.00553751	-7.2112E-06	7.5486E-08	-0.08130617	0.00314976
44.87927014	17.6149169	3035.40791	2914.92511	0.06920345	0.03137966	246.613985	64.4378723	0.65519891	0.5464351	-2.3235E-06	3.8269E-06	0.275381667	0.21253569
12.19368756	2.99881234	370.846189	192.541446	0.00417153	0.00393491	134.326228	12.0311473	0.10494617	0.04774201	-4.269E-06	3.6928E-07	0.121806903	0.02938091
7.763287677	2.66570577	446.005043	323.454551	0.00272089	0.00253339	51.0389769	7.01764317	0.08516787	0.03686051	8.9828E-07	1.9519E-07	0.038142255	0.02728307
0.51516949	0.15702496	3.63489919	1.58451084	0.00010188	7.5307E-05	42.0689411	0.43857269	0.00803354	0.00426077	-6.6327E-06	4.06E-08	0.056216405	0.00109707
38.26448901	18.8396486	2391.92278	2925.66518	0.01681784	0.01649164	254.249262	45.808857	0.39295176	0.24101748	-2.0191E-06	1.2598E-06	0.247781872	0.18669553
15.19947274	3.26995239	397.503096	188.483897	0.01056143	0.00599817	160.28478	17.908461	0.18921102	0.11620722	-3.9622E-06	4.4816E-07	0.122924382	0.07133996
6.765248223	2.5056383	326.975361	245.770791	0.01391056	0.00666774	39.4324236	9.9163013	0.10622501	0.11477616	5.6595E-07	3.8214E-07	0.06332295	0.04128548
0.840555105	0.15743577	2.83530078	0.33482707	8.5554E-05	3.5601E-05	25.8344129	0.81627127	0.00644259	0.00429114	-6.5105E-06	3.873E-08	-0.13790239	0.00513595
36.86312049	19.2324681	2288.29101	4062.88377	0.13160086	0.05101626	248.431178	67.5753264	0.55763407	1.03408786	-2.4154E-06	3.6222E-06	0.416945948	0.32984
12.95814404	2.68983207	323.954211	169.087935	0.01714326	0.00797725	149.692766	17.0148195	0.18584945	0.10531416	-4.2209E-06	4.2739E-07	0.102527581	0.06750318
7.43556807	2.94986407	335.939162	286.183702	0.02344132	0.00870932	46.6309487	11.0751	0.12642613	0.08757272	7.9391E-07	3.3519E-07	0.05988429	0.06493028
-0.17165533	0.10127902	1.38136232	0.08705312	0.0003372	0.00011395	13.5992147	0.84701751	0.00722867	0.00520651	-6.9078E-06	4.4564E-08	-0.23823768	0.00535747
37.71628333	16.5137947	1831.84297	2683.11338	0.1359496	0.05169591	251.209479	65.3670763	0.65924806	0.58767625	-2.0851E-06	2.7675E-06	0.389075235	0.73990512
16.81427166	2.62453021	480.098049	172.243884	0.00389019	0.00283863	184.105706	8.61989082	0.09374048	0.04674242	-3.7346E-06	2.344E-07	0.140313946	0.03405323
4.928193547	2.42633381	326.070156	217.226491	0.00262749	0.0016256	39.812581	6.45576778	0.08690903	0.03946365	5.6275E-07	1.5136E-07	0.043816833	0.02754757
7.837331634	0.212101	75.1892796	7.45597944	5.4477E-05	2.7479E-05	91.2716185	0.44619501	0.00312251	0.00186811	-4.8398E-06	2.797E-08	0.078954359	0.00495739
37.41754786	10.5546641	2383.95488	994.544625	0.01386413	0.01144638	241.731807	29.4881431	0.3720718	0.17543349	-2.479E-06	1.8676E-06	0.302779414	0.12841165

12.2. HISTORY OF ELECTROGLOTTOGRAPHY – TABLE

Table 12.2: History of Electroglottography.

AUTHORS	POPULATION	SAMPLE SIZE	GENDER	FINDINGS
Fabre, 1940.	Unaffected by vocal tract pathologies.	N/A	N/A	Bioimpedance measurements were collected from tracheal level using electrodes as a method proposed for registration of arterial pulse frequencies.
Fabre, 1957.	Unaffected by vocal tract pathologies.	N/A	N/A	'High frequency glottography' was used in studies of human phonation and vocal cord function.
Chevrie-Muller, 1964.	Affected by vocal tract pathologies.	N/A	N/A	Investigated the use of electroglottography in diagnosis of specific disorders, such as stuttering.
Fant *et al.*, 1966.	Affected by vocal tract pathologies.	Several	Male	Validated the EGG waveform against voice inverse filtering method.
van Michel, 1967.	Affected by vocal tract pathologies.	6	Male and Female	Presented waveform patterns of a hypokinetic voice disorder, hyperkinetic voice disorders including recurrent paralysis, nodule, abduction hypotonicity, and ventricular phonation, as well as subjects unaffected by voice pathologies.
Frokjaer-Jensen and Thorvaldsen, 1968.	N/A	N/A	N/A	Presented the electrical circuit of an electroglottograph based on Fabre's design.
van Michel and Raskin, 1969.	Unaffected by vocal tract pathologies.	N/A	N/A	Developed an electroglottograph 'Mark 4 EGG'.
van Michel *et al.*, 1970.	Unaffected by vocal tract pathologies.	1	Male	Concluded EGG signal can be correlated with opening and closing of vocal cords based on comparison between EGG signals and simultaneously captured high speed films.
Fourcin and Abberton, 1971.	Affected by vocal tract pathologies.	Several	Male and Female	Showed EGG signals (here referred to as Lx waveform) can be used for voice quality evaluation – proved that EGG signal varies depending on voice qualities; presented different waveforms for normal, breathy, and creaky voice. First application of a laryngograph – based on a pair of double electrodes, with ground reference.
Abberton and Fourcin, 1972.	Affected by vocal tract pathologies.	Several	Male and Female	Enhanced previous work showing EGG signal waveforms of unilateral paralysis, laryngitis, hoarse voice, and a deaf speaker. Explained how EGG can be used to extract human voice fundamental frequency, suggesting for the purposes of fundamental frequency extraction the EGG signal is simpler than acoustic signal.
Lecluse *et al.*, 1975.	Unaffected by vocal tract pathologies.	1	N/A	Appropriately represented the opened / closed phase, where highest amplitude of EGG signal corresponds to the lowest impedance measurement, meaning closed phase (**Y-axis corresponding to value of vocal fold contact area**). One of EGG model used on excised larynx exhibited responses to acoustic vibrations and demonstrated variations in waveforms across different vowels. The

				remaining instruments provided indications of vocal fold contact.
Wechsler, 1976.	Affected by vocal tract pathologies.	20	Male and Female	Noted differences in frequency distribution in patients experiencing vocal tract pathologies before and after voice therapy. Argues that EGG can detect anomalous laryngeal function even when voice appears normal.
Pederson, 1977.	Unaffected by vocal tract pathologies.	20	Male and Female	Confirmed EGG signal can be correlated with opening and closing of vocal cords based on comparison between EGG signals and stroboscope. Defined sequential stages of the opening and closing of vocal cords. Concluded electroglottography may be useful in medical applications.
Rothenberg, 1981.	Unaffected by vocal tract pathologies.	3	Male and Female	Gave insight into two parameters recordable using EGG: air flow at the glottis and the vocal fold contact area (VFCA). Representation of EGG signal with **Y-axis corresponding to value of impedance**.
Smith, 1981.	Unaffected by vocal tract pathologies.	N/A	N/A	Argued EGG is unreliable as a medical tool due to signal being influenced by acoustic vibrations of the larynx.
Hanson *et al.*, 1983.	Affected by vocal tract pathologies.	4	Male and Female	Validated EGG with photoglottograms, computed open and speed quotients. Concluded open quotient differs between a patient unaffected by voice pathologies, Parkinson's disease, spastic dysphonia, and arsenic poisoning.
Smith and Childers, 1983.	Affected by vocal tract pathologies.	24	Male and Female	Provided results that EGG signals with the application of discriminant analysis can distinguish speakers with pathological larynges from those with larynges unaffected by pathologies with 75% accuracy. Concluded EGG may be useful in medical applications. EGG signal represented in its inverted form, where highest amplitude of EGG signal corresponds to the lowest contact area between vocal cords, meaning highest impedance measurement (**Y-axis corresponding to value of impedance**).
Childers and Larar, 1984. Childers and Krishnamurthy, 1985.	Affected by vocal tract pathologies. Unaffected by vocal tract pathologies.	Several	Male and Female	Argued instants of glottal closure and opening can be identified from EGG by using EGG and simultaneous high-speed films. Suggested the use of EGG derivative as a meaningful parameter for medical assessment of vocal fold physiology. Explored the concepts of closed and open quotients thoroughly explaining their characteristics. Suggested EGG can assist with discrimination of pathological larynx from larynx unaffected by pathologies due to its abnormal vibratory patters. Represented the EGG signal it in its inverted form, where highest amplitude of EGG signal corresponds to the lowest contact area between vocal cords, meaning highest impedance measurement (**Y-axis corresponding to value of impedance**).
Rothenberg and Mahshie, 1988.	Unaffected by vocal tract pathologies.	5	Male and Female	Described a method for estimating the degree of vocal fold abduction from EGG signal (**Y-axis corresponding to value of impedance**), based on a threshold method – chosen level line based on percentage of the amplitude between its minimum and maximum within one glottal cycle (50% for a normal to pressed voice and 35% for a relaxed voice). Found the method in robust, but by its nature it is imprecise and should be interpreted with care.

Titze, 1990.	Unaffected by vocal tract pathologies.	8	Male and Female	Proved crucial influence of electrode size and orientation on the signal-to-noise ratio and linearity of the EGG signal. Suggested better results are obtained in small inter-electrode distance and electrode angle.
Childers *et al.*, 1990.	Affected by vocal tract pathologies.	12	Male and Female	Formulated mathematical equation representing a mathematical model of a EGG waveform. Suggested the use of EGG derivative as a meaningful parameter for medical assessment of vocal fold physiology. Concluded certain EGG features can be associated with vibratory characteristics of both pathological and not-pathological larynges. Represented the EGG signal it in its inverted form, where highest amplitude of EGG signal corresponds to the lowest contact area between vocal cords, meaning highest impedance measurement (**Y-axis corresponding to value of impedance**).
Colton and Conture, 1990.	SYSTEMATIC REVIEW	N/A	N/A	Identified and organised the pitfalls of electroglottography, including easily distorted nature of the measurements, difficulty in electrode placement, electrode-to-skin ratio influencing the measurements, as well as differences between recordings obtained from children, male, and female subjects. Showed that the presence of mucus affects the EGG signal. Confirmed the advantages of EGG, listing the accurate duty cycle, fundamental frequency acquisition remaining more accurate than its extraction from acoustic signals, as well as accurate closing time representation (**Y-axis corresponding to value of impedance**). Concluded the identification of longer closing time in EGG can constitute to accurate disclosure of illnesses such oedema, nodules or tumours.
Kitzing, 1990.	SYSTEMATIC REVIEW with consultations with 17 specialists in the field	N/A	N/A	Argued the use of EGG as a sole diagnostic tool in unreliable, but in conjunction with other methods as photoglottography or stroboscopy it provides valuable additional medical information unobtainable with other methods. Concluded EGG is the best method for measurement of glottal vibratory period, as well as the quotients.
Childers and Lee, 1991.	Unaffected by vocal tract pathologies.	52 healthy, 23 pathological	Male and Female	Used EGG to differentiate four voice types (modal, vocal fry, falsetto, and breathy) through pulse width, pulse skewness, the abruptness of glottal closure, and turbulent noise. Suggested the results of voice investigation with the application of EGG can be used for healthy vs. pathological voice modelling.
Rothenberg, 1992.	N/A	N/A	N/A	Developed a multichannel EGG, allowing for more pairs of electrodes to be connected. Developed alternative electrode configuration for EGG, where single electrodes can be connected either in parallel or in series. Representation of EGG signal with **Y-axis corresponding to value of impedance**.
Baken, 1992.	SYSTEMATIC REVIEW	N/A	N/A	Confirmed EGG signal is an ideal mean for fundamental frequency measurement, and that is it free of supraglottal influence or other variables, such as the airflow, thus disagreeing with Smith [36]. Acknowledged two different representations of EGG waveform in relation to Y-axis implication, suggested that most appropriate representation of EGG signal is with **Y-axis corresponding to value of vocal fold contact area**.

Logemann, 1994.	Unaffected by vocal tract pathologies.	N/A	N/A	Suggested EGG is a successful tool in study of swallowing, non-invasive alternative to videofluorographic imaging.
Hillman *et al.*, 1997.	N/A	N/A	N/A	Suggested EGG is a reliable clinical tool for medical diagnostic while used along videostroboscopic assessment.
Laukkanen *et al.*, 1999.	Unaffected by vocal tract pathologies.	2	Male and Female	Compared Rothenberg's dual-channel EGG [20] with videofluoroscopy, confirming similar trends in larynx's vertical movements, but disagreements in the amount of these movements depending on shifts in the larynx's initial position and changes in the position of cartilages. Suggested multichannel EGG is valid in clinical application, but its applicability for studying laryngeal biomechanics is limited.
Carding *et al.*, 1999.	Affected by vocal tract pathologies.	45	Male and female	Found the EGG signal can be assessed qualitatively by clinicians to establish the process of minor non-organic laryngeal pathologies treatment. Suggested EGG signal is a suitable method for medical assessment of those illnesses and larynx function.
Rothenberg, 2002.	N/A	N/A	N/A	Discussed how choice of high-pass filter cut-off frequency can distort the EGG waveform. Proposed hardware and software methods for adequate phase correction.
Zagolski and Carlson, 2002.	Affected by vocal tract pathology.	16 healthy, 22 pathological	Female	Found that EGG is a reliable method for vocal fold paralysis diagnosis. Concluded EGG is a suitable tool for measuring progress during therapy of vocal fold paralysis.
Henrich *et al.*, 2004.	Unaffected by vocal tract pathology.	18	Male and Female	Thoroughly investigated and discussed EGG derivative and glottal instants derived using dEGG signal. Applied a correlation-based algorithm (DECOM – DEgg Correlation-based Open quotient Measurement) to automatically calculate fundamental frequency and open quotient from dEGG. Suggested dEGG peaks are related to instants of glottal opening and closing, however, only for a healthy male voice.
Kob and Frauenrath, 2009.	Unaffected by vocal tract pathology.	N/A	Male and Female	Suggested multichannel EGG with 12 electrodes (36 channel measurements, time-multiplex algorithm) is a reliable tool in clinical application, the diagnosis of voice, speech, and swallowing disorders.
Vertigan *et al.*, 2008.	Affected by vocal tract pathology.	56 chronic cough, 8 paradoxical vocal fold movement (PVFM), 55 combined CC-PVFM, 25 muscle tension dysphonia, 27 healthy	Male and Female	Found that EGG with simultaneous application of audio analysis is a suitable and effective method for chronic cough and paradoxical vocal fold movement assessment. Study based on statistical approach and manual comparison of parameters: mean fundamental frequency, standard deviation of fundamental frequency, jitter and harmonic to noise ratio.
Sarvaiya *et al.*, 2009.	N/A	N/A	N/A	Published details related to EGG circuit design.
Gibson and Vertigan, 2009.	Affected by vocal tract pathology.	50 untreated, 47 treated.	Male and Female	Found that EGG derived fundamental frequency distribution and the duration of the closed phase show no significant changes between participants before and after speech pathology treatment.

Qin *et al.*, 2009.	Unaffected by vocal tract pathology.	1	Female	Used EGG and HSV (high-speed video) integrated system for investigation of vocal fold vibration inverse parameters. Focused on glottal instants based on dEGG signal. Concluded the integrated system was more accurate than usual methods for inverse parameters of vocal fold vibration.
Thomas and Naylor, 2009.	Unaffected by vocal tract pathology.	N/A	N/A	Proposed SIGMA algorithm for accurate detection of glottal opening and closing instants, with the application of multiscale analysis for singularity detection, group delay function for spike detection, and Gaussian mixture modelling for removal of detections with unlikely features. Achieved accuracy of 99.47% for GCI detection, and 99.35% for GOI detection – most accurate results for glottal instants detection up to date.
Herbst *et al.*, 2010.	Unaffected by vocal tract pathologies.	N/A	N/A	Developed “wavegrams” – a highly successful method for analysing and displaying EGG signals and their first derivatives. Wavegram image represents variations in vocal fold contact as a sequence of events changing with pitch, loudness and voice type. It provides insight into individual glottal cycles, time-varying fundamental frequency of EGG signal, and changes of vocal fold contact phase.
Hosokawa *et al.*, 2012.	Affected by vocal tract pathologies.	19 healthy, 19 dysphonic, 19 afftected by muscle tension dysphonia	Male and Female	Found that EGG parameters pertaining to regularity of vocal fold vibration are a valid diagnostic tool for muscle tension dysphonia.
Ayazi *et al.*, 2012.	Affected by vocal tract pathologies.	55 healthy, 32 pathological	Male and Female	Found that Gastroesophageal Reflux patients had significantly higher irregularity in both voice frequency and amplitude based on EGG measurements.
Yamout *et al.*, 2013.	Affected by vocal tract pathologies.	15 healthy, 24 pathological	Male and Female	Found that EGG derived mean closed quotient for sustained vowels [a] and [e] in multiple sclerosis and healthy participants are comparable, except in patients with dysphonia. Suggested EGG is a reliable tool for dysphonia diagnostics, however, it is not for multiple sclerosis recognition.
Herbst *et al.*, 2014.	N/A	N/A	N/A	Suggested that positive and negative dEGG peaks do not necessarily precisely coincide with events of glottal closure and initial opening. Research based on excised canine larynx, time-synchronized EGG, and ultra-HSV.
Tang *et al.*, 2015.	Unaffected by vocal tract pathologies.	N/A	Male and Female	Proposed that the utilisation of EGG electrodes positioned at an angle, to be employed concurrently with an ultrasound measurement probe placed directly on the larynx, could yield sufficient EGG waveforms encompassing significant reference points indicative of the utmost augmentation and reduction in vocal fold contact area (VFCA).
Barona-Lleo and Fernandez, 2016.	Unaffected by vocal tract pathology.	44 children with ADHD, 35 non-affected children	Male and Female	Showed that children with ADHD suffer significantly more often from dysphonia or hyperfunctional vocal behaviour as compared to children unaffected by ADHD. With the application of audio, EGG and endoscope, found that over 78% of ADHD affected children suffer from vocal nodules.

Somanath and Mau, 2016.	Affected by vocal tract pathologies.	12 healthy, 12 pathological	Male and Female	Built a digital spasmodic dysphonia detection system based on EGG signals. Concluded EGG is unable to differentiate signals gathered from affected and unaffected by an illness participants.
Hampala *et al.*, 2016.	N/A	N/A	N/A	Investigated relation between EGG and actual vocal fold contact area, concluded EGG deviates slightly from VFCA, and although can be a reasonable first approximation, but its results must be interpreted with caution. Research based on deer larynges.
Borsky *et al.*, 2017.	Unaffected by vocal tract pathologies.	11	Male and Female	Classified modal, breathy, rough, pressed, and soft voice types based on EGG signal, using MFCCs as feature extraction method, and cepstral-based features and multivariate Gaussian mixture model for classification. Achieved 83% frame-level accuracy and 91% utterance-level accuracy. Argued different voice types can be classified using MFCCs due to differences in frequency content.
Syndergaard *et al.*, 2017.	N/A	N/A	N/A	Proposed a method for VFCA vs EGG signal investigation by creating electrically conductive vocal fold replicas.
Ramirez *et al.*, 2017.	Affected by vocal tract pathologies.	17 healthy, 17 pathological	Male and Female	Using EGG and audio analysis, established that shimmer, jitter, open quotient, and irregularity are significantly increased in the patients with Laryngopharyngeal Reflux.
Szklanny *et al.*, 2019.	Affected by vocal tract pathologies.	37 healthy children, 57 affected by vocal fold nodules.	Male and Female	Found that in evaluation of vocal fold nodules in children the EGG signals are far more accurate – changes in EGG were detected in 95% of children with vocal fold nodules, while acoustic signals only confirmed the 63% of affected children. Investigated EGG through Closed Quotient, and audio through Peak Slope – calculations computed and evaluated manually.