Audio-based deep learning classification of laryngeal pathologies with detection of precancerous and cancerous lesions using Gammatone Cepstral coefficients

Biomedical Engineering Advances 11 (2026) 100211

Contents lists available at ScienceDirect

Biomedical Engineering Advances

journal homepage: www.journals.elsevier.com/biomedical-engineering-advances

Audio-based deep learning classification of laryngeal pathologies with detection of precancerous and cancerous lesions using Gammatone Cepstral coefficients

Julia Zofia Tomaszewska [a,*], Wojciech Kukwa [b], Apostolos Georgakis [a]

[a] School of Computing and Engineering, University of West London, London W5 5RF, UK
[b] Department of Otorhinolaryngology, Faculty of Medicine and Dentistry, Medical University of Warsaw, 02-091 Warsaw, Poland

ARTICLE INFO

Keywords:
Laryngeal pathology
Gammatone Cepstral coefficients
Voice pathology classification
Voice pathology detection
Deep learning
Convolutional neural networks
Audio signal classification
t-test

ABSTRACT

Introduction: Despite extensive research on audio-based voice pathology detection, current literature lacks clear and consistent evidence identifying acoustic features capable of reliably discriminating precancerous and cancerous laryngeal lesions, particularly when analysed using continuous speech signals.
Problem statement: The performance of audio-based laryngeal pathology classification systems on continuous speech remains significantly underreported, and commonly used Mel-Frequency Cepstral Coefficients (MFCCs) may be suboptimal for capturing pathology-related acoustic characteristics.
Objectives: This study investigates the hypothesis that continuous speech audio signals analysed with Gammatone Cepstral Coefficients (GTCCs) enable the accurate and precise detection of laryngeal pathologies, with the specific focus on precancerous and cancerous lesions.
Methods: An audio-based classification system employing GTCCs for feature extraction and a one-dimensional Convolutional Neural Network (CNN) for classification is proposed. The system considers three classes: precancerous and cancerous lesions, neuromuscular disorders, and healthy cases. Performance was evaluated using two datasets: a custom speech dataset collected for this research and the Saarbruecken Voice Database (SVD).
Results: GTCCs derived from speech signals delivered superior classification accuracy compared to the widely used Mel-Frequency Cepstral Coefficients (MFCCs). On the custom dataset, the proposed method achieved an average classification accuracy of 85.04% ±1.23 compared to 63.22% ± 1.62 using MFCCs. On SVD, GTCCs achieved 73.93% ±1.42, compared to 60.36% ±2.44 for MFCCs. The statistical significance of the obtained results was evidenced using *t*-test with the significance level set at 1%.
Conclusions: The results demonstrate that GTCCs extracted from continuous speech signals provide a robust and effective representation for audio-based laryngeal pathology classification, highlighting their potential for use in automated pre-screening systems targeting precancerous and cancerous voice disorders.

Introduction

Laryngeal pathologies cover a range of medical conditions affecting the organ responsible for voice production, breathing, and protection of the airways during swallowing. Varying in severity from a range of inflammations and benign nodules to malignant lesions, those pathologies have a substantial lifetime prevalence estimated up to 29.9 % [1]. The accurate and timely identification of those diseases is vital for an effective treatment. Current diagnostic methods, such as laryngoscopy and stroboscopy, involve the insertion of an endoscope into the throat for direct visualisation, followed by the subsequent image interpretation by a specialist. Therefore, these methods are invasive, subjective, and not ubiquitously available [2], highlighting the need for an assistive, rapid and dependable classification system for laryngeal pathologies.

In response to the growing demand for an automated vocal tract pathology screening tool, recent efforts have focused on developing systems to detect pathologies using audio recordings of vocal sounds obtained from affected individuals [3–15]. Current research has mostly considered binary classification between healthy and pathological signals [3–7,9,11] or between two broad categories of vocal tract conditions, often focusing on the structural differences [14,15]. Many of those methods rely on the Mel-frequency spectrum, particularly

[*] Corresponding author.
E-mail address: julia.tomaszewska@uwl.ac.uk (J.Z. Tomaszewska).

https://doi.org/10.1016/j.bea.2026.100211
Received 13 November 2025; Received in revised form 13 January 2026; Accepted 18 January 2026
Available online 19 January 2026

Mel-frequency Cepstral Coefficients (MFCCs) for their feature extraction stage [3,7,8,12,13,17]. Furthermore, significant majority of the currently existing laryngeal pathology classification systems relies on audio recordings of sustained vowel phonation, especially vowel /a/ [3, 6–12,14–17]. There is a considerable lack of reporting on the effectiveness of designed laryngeal pathology classification systems on continuous speech signals instead of sustained phonation.

Despite the growing body of work in this area, limited success has been reported in the classification of a wider range of specific laryngeal conditions using audio recordings. This may partly stem from the application of unsuitable feature extraction methods. For instance, features derived from the Mel spectrum may not be capable of adequately capturing those characteristics essential for the accurate discrimination between various laryngeal pathologies. In that case, the use of an alternative, better-suited feature extraction approach could improve the performance of an audio-based laryngeal pathology classification system.

In the following work, we present an audio-based classification system, capable of accurately discriminating between three laryngeal pathology groups: precancerous and cancerous lesions (class I), neuromuscular conditions (class II), and healthy cases (class III). The proposed system relies on Gammatone Cepstral Coefficients (GTCC) for the feature extraction method, and one-dimensional Convolutional Neural Network (CNN) for the classifier. Due to the significant lack of reporting on performance of laryngeal pathology classifiers on continuous speech instead of sustained vowel phonation, the system proposed in this work was trained and validated using audio recordings of speech. For generalisability purposes, two different datasets were used; a dataset of speech audio recordings developed for the purposes of this research, as well as the commonly used Saarbruecken Voice Database (SVD).

This work aims to:

1) Propose a novel laryngeal pathology classification system capable of detecting precancerous and cancerous laryngeal lesions, while also distinguishing them from non-cancerous conditions, in a 3-class scenario.
2) Demonstrate that GTCCs are superior to MFCCs in classification of laryngeal pathologies based on audio recordings.
3) Report on the performance of continuous speech instead of sustained phonation in a laryngeal pathology classification system, demonstrating how this choice can enhance the detection accuracy of cancerous and precancerous laryngeal lesions.
4) Introduce a new dataset of audio recordings of speech, gathered from individuals affected by various laryngeal pathologies, including, among others, laryngeal lesions, neuromuscular conditions, as well as a control group of healthy participants.

Unlike previous studies that primarily relied on sustained vowel phonation and Mel-frequency features, this work demonstrates that continuous speech retains discriminative information essential for identifying subtle pathological changes. This approach provides a more realistic basis for developing automated voice-based screening tools deployable in everyday clinical and digital environments.

The remainder of this paper is organised as follows: Section 2 introduces the theoretical background of the ERB spectrum and GTCC feature extraction while comparing them to Mel spectrum derived features; Section 3 reviews related work; Section 4 describes the datasets, feature extraction process, and CNN architecture; Section 5 presents the experimental results and statistical analysis; and Section 6 discusses the findings, limitations, and future research directions.

Preliminaries

The Equivalent Rectangular Bandwidth (ERB) spectrum provides an insight into the human auditory perception by modelling the frequency bandwidths of natural auditory filters present in the cochlea [18]. Unlike the Mel-frequency spectrum, which relies on subjective perception [19], the ERB spectrum aligns with the human auditory system's physiological mechanisms. To represent the ERB spectrum, the Gammatone filter bank is utilised, simulating the frequency decomposition occurring at the cochlea [18]. The filter bank is referred to as "gammatone" since the impulse responses $g(t)$ is obtained as the product of multiplication between the gammatone distribution function and a sinusoidal tone:

$$g(t) = A \cdot t^{n-1} \cdot e^{-2\pi Bt} \cdot \cos(2\pi f_c t + \varphi)$$

where A is a gain constant (usually equal to 1), t represents time, n is the order of the filter, B is the bandwidth of the filter, f_c is the centre frequency of the filter, and φ is the initial phase shift of the filter [20].

Gammatone Cepstral Coefficients (GTCCs) are derived from the ERB spectrum, offering a representation of sound features based on the ERB frequency bands. They capture the nonlinear frequency response of the human auditory system [22], serving as the biologically inspired counterparts to Mel-frequency Cepstral Coefficients (MFCCs), which are instead derived from the Mel spectrum [14,21]. The process of deriving GTCCs involves similar steps to MFCC derivation, starting with Fourier Transform calculation, application of the appropriate filter bank (Mel filter bank for MFCCs and Gammatone filter bank for GTCCs), logarithmic compression, and performing of the Discrete Cosine Transform [22]. Fig. 1 shows the corresponding frequency scale characteristics of the Mel and Gammatone filter banks. MFCCs are characterised with coarse triangular filter responses (Fig. 1a), resulting in limited overlap between adjacent filters. In contrast, GTCCs utilise smoother filter responses (Fig. 1b), enhancing the overlap between filters and minimising the spectral information loss. Furthermore, the ERB scale results in a higher filter bank resolution, especially at lower frequencies, compared to the Mel scale.

From a physiological perspective, the ERB-based representation underlying GTCCs is better suited for the analysis of pathological voice signals in comparison to Mel-frequency representations. Laryngeal pathologies, especially precancerous and cancerous lesions, introduce irregularities in vocal fold vibration, altered harmonic structure, and changes in energy distribution across the frequency bands, which result from the presence of lesions, stiffness asymmetries, and incomplete glottal closure. Compared to Mel spectrum, ERB scale provides higher frequency resolution at lower frequencies and smoother filter overlap, possibly enabling more precise capture of subtle spectral variations associated with pathological phonation. When applied to continuous speech, these properties allow GTCCs to represent both sustained phonatory characteristics and dynamic articulatory transitions, which may carry additional information related to pathological voice production. Furthermore, ERB spectrum presents a biologically motivated spectral representation, in contrast to Mel spectrum based primarily on perceptual scaling.

Previous work

Research on audio-based voice pathology detection has primarily relied on statistical approaches and machine learning techniques [4,5,7, 8,10]. These methods have been used to process a wide range of features, such as pitch perturbation, amplitude perturbation, harmonic-to-noise ratio [3–5,7,9], as well as Mel spectrum derived features [3,7,13]. Methods using Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) have been particularly successful in distinguishing between healthy and pathological audio signals, achieving accuracies of 98.1 % [5] and 98.23 % [7]. Work relying on acoustic features of reduced dimensionality using Principal Component Analysis (PCA), and Random Forest classifiers, have further reported accuracies of up to 100 % [9]. Deep learning methods have also been proven effective in binary classification between control and pathological audio signals [3,6,11].

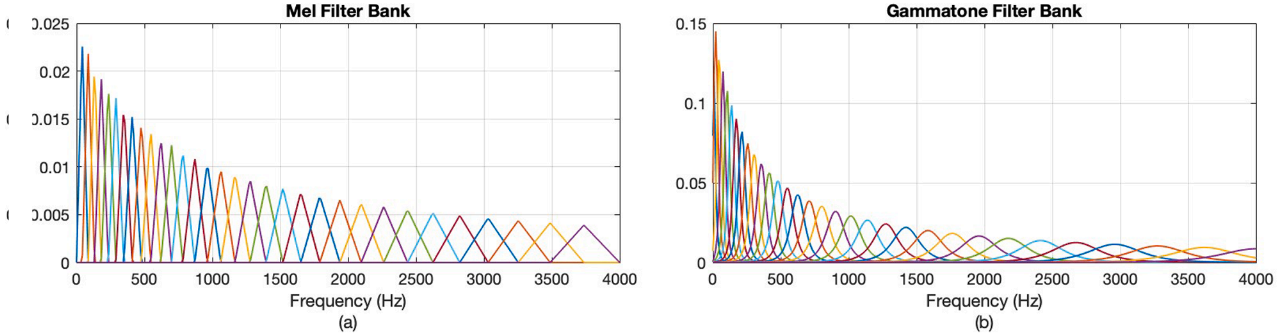

Fig. 1. a) Visual representation of a Mel filter bank. b) visual representation of the Gammatone filter bank.

In more recent studies, deep learning methods have further been successfully applied in classifying laryngeal pathologies based on aetiological distinctions, e.g. structural and neuromuscular lesions [13, 15]. Those systems have relied on Bi-Directional Long-Short Term Memory Networks (BiLSTM), reporting 89.27 % accuracy [13], as well as Convolutional Neural Networks (CNN) achieving the accuracy between 72 % and 88 % [15].

The investigation into alternative feature extraction methods, such as ERB spectrum-derived features [14,21,23] has also led to successful results. The system proposed in [14] implemented Gammatone Spectral Latitude as the feature extraction method and Multilayer Perception (MLP), Support Vector Machine (SVM) and Random Forest (RF) as the classification methods. The discrimination accuracy between variants of certain pathologies was reported to reach 99.6 %. Nevertheless, in cross-database testing, the achieved classification accuracy was around 70 %, dropping to 55.7 % when applied to SVD. The computed average accuracy of detection of “structural” conditions – which included some of SVD's precancerous and cancerous categories [14] – was equal to 59.61 %. Furthermore, the study considered solely the use of audio recordings of sustained phonation, with no investigation into the use of continuous speech signals.

Research on multiclass discrimination of individual vocal tract disorders is currently limited. The successful “multiclass” laryngeal pathology classifiers rely mostly on binary comparisons between pairs of individual conditions [8,10], whereas true multiclass approaches have reported accuracy results below 75 % [12,17]. Integrating audio data with other modalities, such as medical data or electroglottographic measurements [23], could offer a more reliable laryngeal pathology classification; however, more work is required in this area of research.

It is crucial to note that most research in laryngeal pathology classification has been conducted using sustained phonation [3,6–12, 14–17], with very few studies exploring the use of continuous speech. 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 [6,8,17]. However, continuous speech involves dynamic changes in glottal positioning, which could reveal additional pathological patterns not observable in sustained phonation. Furthermore, there is a noticeable gap in the literature regarding the evaluation of laryngeal pathology classification systems using continuous speech data. In this paper, we address this gap by examining the proposed laryngeal pathology classification system with audio recordings of continuous speech.

Table 1 provides a summary of selected previous work on audio-based classification of vocal tract pathologies.

Methods

The processing steps of the implemented system are depicted in Fig. 2.

Dataset

To ensure the generalisability of this study, the proposed system was trained and tested on two different databases; the custom dataset created specifically for this research, and the publicly available Saarbruecken Voice Database (SVD) [10,11,14,25]. The publication of the dataset created for this project is underway.

Saarbruecken voice database

SVD contains recordings collected from healthy participants, as well as those suffering from one or more of 71 various vocal tract conditions [26]. It consists of audio recordings of sustained vowels “i”, “a”, “u” produced at normal, high, and low pitch, the sustained vowels of rising-falling pitch, as well as speech recordings of utterance “Guten Morgen, wie geht es Ihnen?”. All recordings contained in SVD were conducted at a sampling rate of 50 kHz and a 16-bit depth resolution.

Although SVD contains over 2000 recordings [14,25,26], from which 1356 are pathological [11], it should be noted that this number does not correspond to unique participants. Instead, certain participants have been recorded multiple times – in some cases, as many as 24 times (e.g. participant no 2027, suffering from spasmodic dysphonia), and each recording has been assigned a unique identification number (IDN). In total, 336 recordings under various IDNs belong to the same participants throughout the database. 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 rates. The actual number of unique participants within the pathological subset of SVD is 1020.

The presence of multiple recordings collected from identical subjects is an important point where the SVD user must exercise caution. In particular, to maintain participant independence between the training and validation stages, the user must ensure that recordings belonging to the same subject should not appear in both the training and validation datasets. Such an error would introduce speaker-dependant bias in the classification process, leading to artificially inflated performance validation results. It would be recommended that work using SVD should provide clarification as to the participant-independence of the reported models. In this work, we have assigned unique identification numbers to each participant, and ensured that same-subject recordings were exclusively allocated either in the training or the validation dataset. It

Table 1
Summary of seminal published work on audio-based laryngeal pathology classification.

REFERENCE	DATASET	OBJECTIVE	METHODS	FINDINGS
Moran et al., 2006 [4].	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 /ah/. 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 [5].	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 /ah/. FEATURES: Acoustic features (Multidimensional Voice Program – MDVP), and Principal Component Analysis (PCA). CLASSIFICATION: Support Vector Machine (SVM).	ACCURACY: 98.20 %
Henríquez et al., 2009 [6]	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 /ah/. 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-Londono et al., 2010 [7].	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 /ah/. 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 [8].	226 speakers (173 pathological – 20 “nodules”, 20 “polyps”, 26 “keratosis”, 22 “adductor”, 71 “paralysis”, remaining 14 unspecified, and 53 control) Male and Female.	Binary classification between control and pathological, and binary pairwise discrimination between individual pathologies (“nodules”, “polyps”, “keratosis”, “adductor”), as well as binary pairwise discrimination between “nodules”, “polyps”, “keratosis”, and “adductor” collectively against “paralysis”.	DATA: Massachusetts Eye and Ear Infirmary (MEEI) – sustained phonation vowels /ah/. FEATURES: Modulation Spectral Features, compared to MFCCs. CLASSIFIER: Support Vector Machine (SVM).	Modulation Spectral Features outperforming the MFCCs. 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 [9].	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:	ACCURACY: 99 % accuracy.

(continued on next page)

Table 1 (*continued*)

REFERENCE	DATASET	OBJECTIVE	METHODS	FINDINGS
			Acoustic features, and Principal Component Analysis (PCA). CLASSIFIER: Random Forest Classifier.	
Al-Nasheri et al., 2017 [10].	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 /ah/. 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 [11].	1166 speakers (583 pathological, 583 control). Male and Female.	Binary classification between control and pathological.	DATA: Saarbruecken Voice Database (SVD) – sustained phonation vowels /ah/. 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 [13].	1045 speakers (all pathological – 100 “functional dysphonia”, 103 “neoplasm”, 718 “phonotrauma”, 124 “vocal palsy”). Male and Female.	Multiclass 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: Bi-Directional Long-Short Term Memory Network (BiLSTM).	ACCURACY: 89.27 % maximum.
Zhou et al., 2022 [14].	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.	Multiclass 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 Prı́ncipe 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 %.
Kuo et al., 2023 [15].	523 speakers (415 pathological – 112 “neoplasms”, 303 “benign structural diseases”, and 108 control). Male and Female.	Multiclass 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: 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 %.

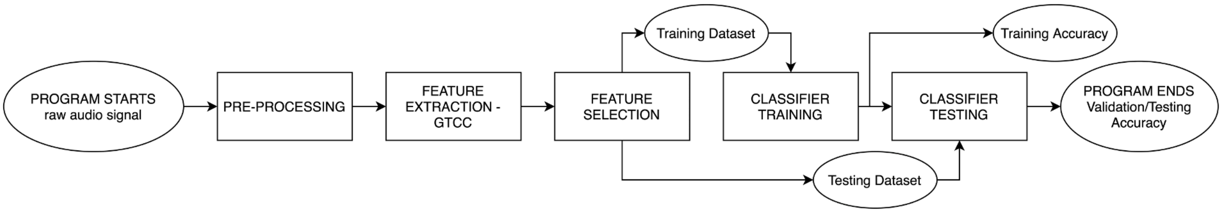

Fig. 2. Block diagram of the processing stages of the system implemented in this study.

was further noted that among the 1020 participants from the pathological group, some belong to conditions that are not typically considered as laryngeal pathologies from a medical perspective. Those included the instances referred within the SVD as:

- “Gesangsstimme” and “Sengerstimme”, denoting the voice of a professional singer, neither being a pathology.
- “Mutatio”, referring to normal voice changes during puberty.
- “Morbus Down”, relating to Down Syndrome of the participant.
- “Morbus Parkinson”, relating to the participants’ Parkinson's Disease.
- “Vox senilis”, referring to natural changes in voice due to aging.

The number of subjects to whom one of the above conditions was assigned as the only condition present was equal to 41. After removing those from the SVD pathological subset, the total number of individual participants within SVD suffering from laryngeal pathologies is equal to 979.

Additionally, it appears that at least 53 participants in SVD’s pathological subset had already undergone treatment before the collection of data, as shown in the medical remarks included in the SVD records (e.g. participants no 1454, 1475, 1743, and others). Those participants have not been involved in this study, as their original pathologies did not fall within the classes considered.

Furthermore, 21 subjects from the 979 pathological subjects were under 18 years of age, which for the purposes of this study were not considered, since we have focused on adult speakers only. This resulted in the pathological subset of SVD totalling 958 individual participants.

Similarly, for the healthy group, while 869 subjects are listed [26], 36 are under 18, and 15 are duplicates (same subjects with different recording IDNs). Therefore, the total number of individual participants in the healthy group of SVD shall be reported equal to 818.

For the purposes of this study, a subset of SVD’s available pathology categories was selected, to align with the aim of classifying among three classes of: precancerous and cancerous lesions, neuromuscular conditions, and healthy subjects. For the sake of reproducibility, below we describe in detail how our classes have been formulated from the categories available in SVD. For the precancerous and cancerous class, the following conditions were chosen from SVD: Hypopharyngeal Tumour (“Hypopharynxtumor”), Laryngeal Tumour (“Kehlkopftumor”), Laryngeal Pachydermia (“Kontaktpachydermie”), Leukoplakia (“Leukoplakia”), Vocal Cord Carcinoma (“Stimmlippenkarzinom”). Some of the cancerous and precancerous pathologies from SVD had to be omitted due the lack of sufficient number of subjects contained within those classes. Those pathologies included Carcinoma in Situ, Epiglottic Carcinoma (“Epiglottiskarzinom”), Mesopharyngeal Tumour (“Mesopharynxtumor”), as well as Papilloma (“Papillom”). These classes offered the data obtained from only one participant each, from which some were recorded post-treatment (e.g. Papilloma and Mesopharyngeal Tumour).

From the total number of participants within the precancerous and cancerous subset of SVD (77), 16 were repetitions of already recorded subjects, and 7 had undergone the surgical treatments, leaving a usable dataset of precancerous and cancerous conditions at the count of 54 participants. Individual counts for specific conditions are presented in Fig. 3, where “A” is Hypopharyngeal Tumour, “B” stands for Laryngeal Tumour, “C” is Laryngeal Pachydermia, “D” – Leukoplakia, and “E” – Vocal Cord Carcinoma.

As for the neuromuscular disorder class, the Recurrent Laryngeal Nerve Palsy (“Rekurrensparese”) condition was chosen. After excluding the repetitions of individual participants, as well as those under the age of 18, the neuromuscular disorder subset remained at 144 unique subjects.

From among 818 healthy participants present in SVD, 200 participants were chosen randomly in each of the cross-validation runs, to reduce the imbalance among the 3 classes.

The final number of unique participants in each category can be seen on Fig. 4.

For the purposes of this research, we used SVD’s recordings of speech utterances only. All recordings were subjected to the same data preprocessing procedures as the recordings from the custom database.

Custom data collection

The dataset for this research consists of audio recordings of continuous speech performed by Polish-speaking participants reading the same paragraph of text in Polish, written specifically for this study. The average duration of this passage across the participants was 26.17±7.25 s (22.32±3.72 for control, 26.67±7.45 for pathological). The same predefined speech passage was used for all participants in both the control and pathological groups to minimise linguistic and content-related variability. The observed differences in recording duration

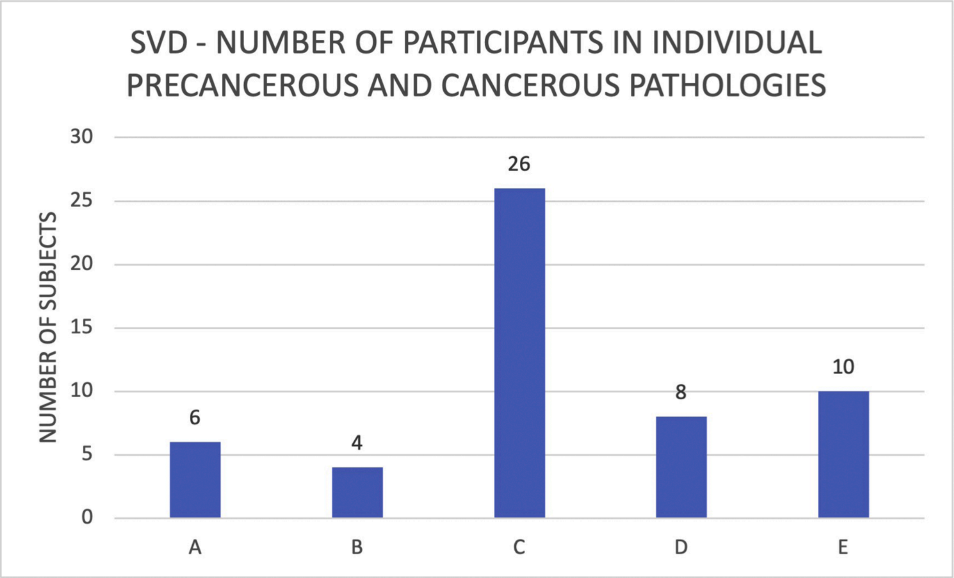

Fig. 3. Unique participants count for specific pathologies chosen from SVD as precancerous and cancerous conditions, where “A” is hypopharyngeal Tumour, “B” stands for Laryngeal Tumour, “C” is Laryngeal Pachydermia, “D” – Leukoplakia, and “E” – Vocal Cord Carcinoma.

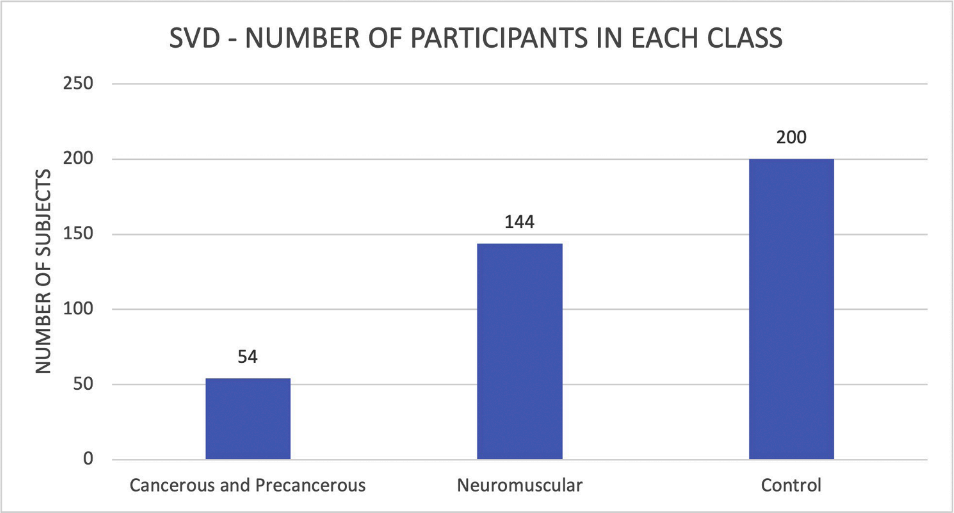

Fig. 4. Number of participants in each category selected from SVD.

reflect natural variations in individual speech rate and articulation rather than any systematic group-dependent design.

The data for this project were collected at the ENT department of Czerniakowski Hospital, Warsaw, Poland, in a dedicated hospital room – all recordings were conducted in the same acoustic conditions. All recordings were carried out in accordance with General Data Protection Regulation, outlined in the *Regulation (EU) 2016/679 (General Data Protection Regulation)*, and ethics approval was obtained.

Immediately before the beginning of the data collection process, all participants were assessed by a clinician specialised in phoniatrics to ensure the correct diagnosis. The total number of participants was 156. This entailed 20 participants in a control group (not affected by any of the laryngeal conditions), as well as 136 subjects suffering from cancerous and precancerous lesions, and neuromuscular or inflammatory laryngeal pathologies, including a variety of disorders, growths affecting vocal folds, growths within the larynx not affecting the vocal folds directly, Reinke's Oedema, neuromuscular conditions including the vocal fold paralysis, as well as functional dysphonia, and laryngitis. The data gathered from 15 participants from pathological group were removed from the overall dataset due to technical issues with the recordings. Furthermore, the number of participants in each category varied, resulting in an unbalanced dataset, which can be considered as one of the limitations of this study.

For the purposes of this research, we selected only the recordings collected from participants affected by precancerous and cancerous conditions, as well as neuromuscular disorders, and the group of healthy subjects as a control group. The final number of participants in each category can be seen in Fig. 5.

The dataset for this project was recorded with the dynamic SM57 microphone and Scarlett 2i4 audio interface. The Logic Pro X digital audio workstation was used, with no compression or other pre-processing methods applied. The recordings were captured with the sample rate of 44.1 kHz and the bit depth of 16 bits per sample. All recordings were extracted from the workstation in a form of mono WAV files.

Data pre-processing

First data pre-processing stage consisted of normalising all signals' amplitude. To avoid any signal alternations to existing recordings, and

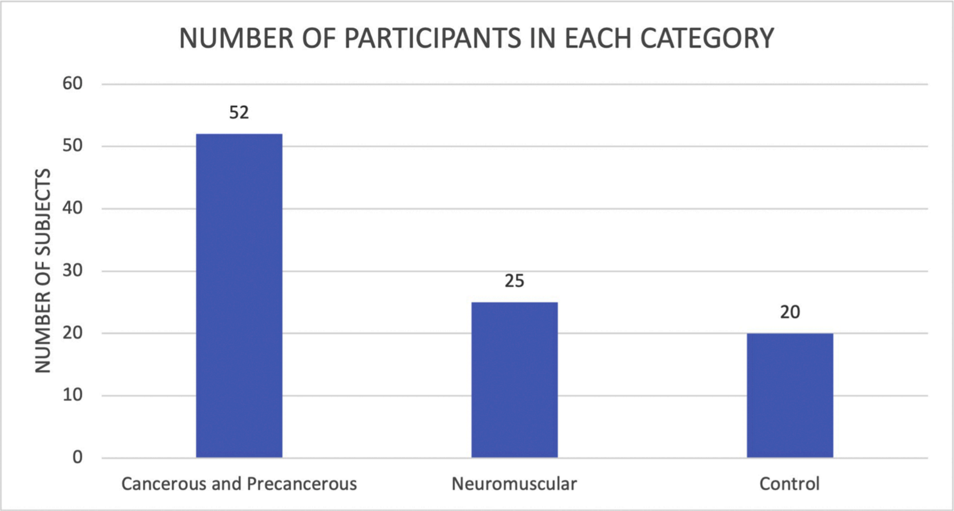

Fig. 5. Number of participants in each category in custom dataset.

Table 2
Number of final data samples used for the classification (after data pre-processing stage).

Class	Custom Dataset	SVD
Cancerous and Precancerous	1361	144
Neuromuscular	792	351
Healthy (Control)	388	771

thus preserve the real information contained within them, 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.

Since the speech recordings varied in length, all data was pre-segmented. To ensure that recording length did not act as a confounding factor in the classification process, all recordings were split in one-second-long files (1 s being within the range of commonly applied time intervals in relevant literature) by saving all subsequent N samples (where $N = \textit{sample rate}$) of one recording as a separate WAV file. This step allowed us to avoid zero-padding. This window length allows sufficient capture of phonatory characteristics while maintaining uniform input dimensions across samples and avoiding zero-padding.

Once fully pre-processed, the custom database consisted of 2541 data samples. The total of samples obtained from SVD was 1266. Table 2 depicts the exact number of data samples in each category.

The distribution of samples across the three classes was uneven, with the pathological categories containing substantially more samples than the healthy control group in the custom dataset, and the opposite pattern observed in SVD. This imbalance was expected due to the natural prevalence of pathological and control recordings within the respective databases. The imbalance was later addressed during the model training stage through an oversampling procedure described in Section 4.3.

Feature extraction

In this study, to ensure the capture of all most relevant features, 40 coefficients were calculated for both MFCCs and GTCCs. The signals were analysed using a sliding window, with an overlap set to half of the frame size. The Hann window was chosen, with a frame size set to 512 samples. Mathematically, the coefficients of Hann window – for a sequence of N length – can be represented as:

$$w(n) = 0.5\left(1 - \cos\left(2\pi \frac{n}{N}\right)\right),\ \textit{for}\ 0 \le n \le N$$

CNN classifier

The architecture of the designed CNN can be seen on Fig. 6.

The proposed CNN architecture is structured around one-dimensional convolutional layers, tailored to accommodate the format of the cepstral features fed into the network. One-dimensional convolutions were employed to preserve the temporal structure of the GTCC feature sequences while maintaining computational efficiency. The network comprises four blocks, each consisting of one-dimensional convolutional layers with a stride of 1, followed by ReLU activation and normalisation layers. The kernel size of the convolutional layers in the first two blocks was set to 3, while in the third and fourth blocks, the kernel size was increased to 10. Smaller kernels capture local spectral–temporal patterns, while larger kernels in deeper layers allow modelling of longer temporal dependencies across speech segments. The number of filters in the convolutional layers is 16 for the first and third blocks, and 32 for the second and fourth blocks. The progressive increase in the number of filters enables the network to learn increasingly complex, higher-level feature representations while controlling model capacity. The initial learning rate was set to 0.01.

To prevent overfitting, dropout layers with a dropout rate of 0.2 were incorporated after the second and fourth blocks. The network also includes a one-dimensional global average pooling layer, followed by two fully connected layers – the first with a size equal to the number of features (40), the second of size equal to the number of classes (3) – separated by a dropout layer with a dropout rate of 0.5. The final layer employs the SoftMax transfer function for classification.

The dataset was fed into the proposed CNN in split of 80 % for training and 20 % for validation (per class). To reduce the impact of class imbalance in both datasets, a simple random oversampling strategy was applied during training. Where one class contained fewer samples than others, its training subset was expanded by duplicating existing data instances until reaching a comparable sample count to the majority class. To ensure the speaker-independence of the proposed model, an additional data segregation algorithm was applied to ensure all samples associated with a specific identification number (corresponding to a particular participant) were exclusively assigned to either the training or the validation dataset. This results in a slight variation in data split ratio between training and validation depending on a cross-validation run. This approach was critical to prevent the classifier from emphasising participant-specific characteristics across datasets, thereby maintaining the integrity of our speaker-independent classification model.

During training, the CNN utilises the Adam optimisation algorithm, which combines the advantages of AdaGrad and RMSProp. This algorithm computes the moving averages of parameter gradients and their squared values, akin to RMSProp, while also incorporating momentum. By adapting the learning rates for each parameter based on their past gradients and squared gradients, Adam enables faster convergence. Thus, Adam was selected due to its robustness and adaptive learning-rate properties when training deep networks on moderately sized datasets.

The proposed CNN system is configured with a mini-batch size of 16 selected to provide stable gradient estimates while maintaining sufficient stochasticity during training. With data shuffling at every epoch, the maximum epoch number was set to 30 to avoid overfitting, as empirically observed during preliminary experiments. The validation frequency set to 64 iterations. All training hyperparameters, including mini-batch size, number of epochs, and validation frequency, were chosen empirically to balance convergence stability, generalisation performance, and computational efficiency.

Results

To further enhance the generalisability of the proposed classification system, a 5-fold cross-validation was implemented; the process of training and validation of the model was repeated five times, each time on a different subset of the available data. To ensure speaker independence, the participants, rather than individual speech segments, were assigned to folds. This ensured that all speech samples originating from a given participant were contained exclusively within either the training or the validation set. The choice of five folds was made to balance reliable estimation of generalisation performance and the limited number of participants, particularly in the pathological classes. This allowed for each participant to contribute to both training and

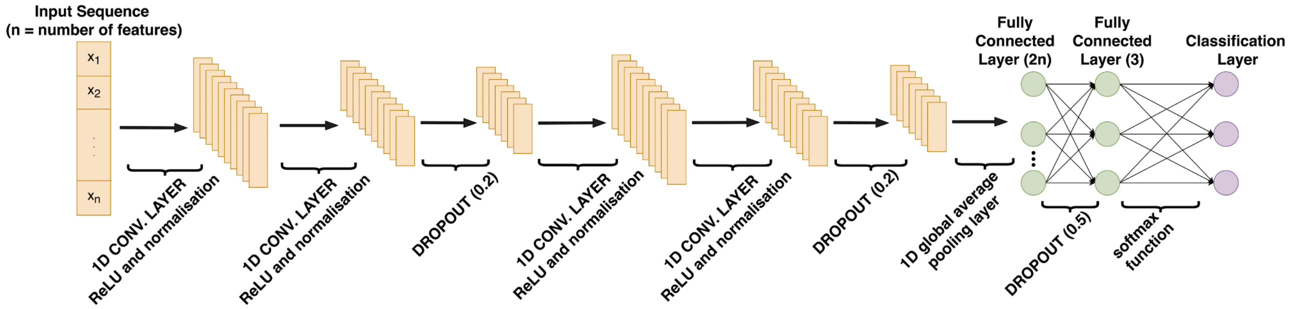

Fig. 6. Architecture of the designed CNN classifier.

validation across different folds while maintaining strict subject independence.

Classification performance metrics were computed for each fold independently and subsequently averaged across folds to obtain the reported results. The statistical significance of the obtained results was evaluated using *t*-test calculation.

All experiments were conducted on a standard desktop workstation. Model training and evaluation were performed using MATLAB, running on a system equipped with a Dual-Core Intel Core i7 processor, 16 GB RAM, and an NVIDIA GPU. The reported results correspond to the average performance across cross-validation folds, with identical training and evaluation procedures applied across all experiments to ensure reproducibility.

The evaluation metrics used to evaluate the achieved classification performance in this study – accuracy, precision, sensitivity, specificity, and F1-score – are standard performance measures widely adopted in classification tasks and biomedical signal analysis. These parameters are defined as follows:

- Precision, 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, 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, a measure of balance between precision and sensitivity represented by their harmonic mean:

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

- 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:

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

Table 3
The classification accuracy scores achieved for each round of cross-validation.

Cross-validation Round	GTCC features, Custom Database	MFCC features, Custom Database	GTCC features, SVD	MFCC features, SVD
1	85.02 %	64.25 %	73.81 %	58.33 %
2	86.67 %	65.56 %	72.02 %	60.12 %
3	85.00 %	62.22 %	73.21 %	60.71 %
4	83.22 %	62.22 %	75.00 %	64.29 %
5	85.29 %	61.83 %	75.60 %	58.33 %
AVERAGE ACCURACY	**85.04 % ±1.23**	**63.22 % ±1.62**	**73.93 % ±1.42**	**60.36 % ±2.44**

- Accuracy, indicating the proportion of correctly classified samples among all possible instances.

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

CNN multiclass laryngeal pathology classifier

The proposed classifier fed with GTCCs achieved an average accuracy of 85.04 % with a standard deviation of 1.23 on the custom dataset, and an average accuracy of 73.93 % with standard deviation of 1.42 on SVD. When fed with MFCCs derived from the custom dataset, the same classifier achieved an average accuracy of 63.22 % with standard deviation of 1.62. When fed with MFCCs derived from the SVD, the average accuracy was 60.36 % with standard deviation of 2.44. The results demonstrate the superiority of GTCCs over MFCCs as discriminant features from audio signals for laryngeal pathology classification.

Table 3 lists the accuracy scores achieved on both datasets for each round of cross-validation. The reported classification accuracy was calculated by dividing the sum of the positive predictions for all three classes by the total number of instances:

$$Classification\ Accuracy = \frac{TP_c + TP_n + TP_h}{Total\ Nr\ of\ Instances}$$

where TP_c stands for positive predictions of cancerous and precancerous instances, TP_n reflects the number of correctly identified neuromuscular instances, and TP_h represents the positive predictions of healthy

Table 4
Average confusion matrix calculated for the proposed system fed with GTCCs derived from the custom dataset.

GTCC on custom data		PREDICTED CLASS		
		Cancerous and Precancerous	Neuromuscular	Healthy
TRUE CLASS	Cancerous and Precancerous	**80.76 %**	12.93 %	6.31 %
	Neuromuscular	19.87 %	**77.60 %**	2.52 %
	Healthy	0.95 %	2.21 %	**96.85 %**

Table 5
Average precision (Pr), sensitivity (RC), specificity (Sp), and F1-score reported as % achieved by GTCCs derived from the custom dataset.

Parameter	Cancerous and Precancerous	Neuromuscular	Healthy
Precision	80.08 %±5.93	84.62 %±7.32	91.62 %±4.54
Sensitivity	80.43 %±10.09	77.82 %±8.53	96.95 %±3.82
F1-Score	79.78 %±4.96	80.59 %±3.60	94.16 %±3.48
Specificity	89.60 %±4.63	92.50 %±3.79	95.49 %±2.52

instances.

Table 4 represents the average confusion matrix calculated from all cross-validation instances of the proposed method while fed with GTCC derived from the custom dataset. Table 5 denotes the values of Precision, Sensitivity, Specificity, and F1-score parameters for that method, calculated for each pathology class separately. Table 6 and Table 7 present the confusion matrix, as well as the table of Precision, Sensitivity, Specificity, and F1-score parameters calculated for the proposed classifier fed with the MFCCs derived from the custom dataset. Tables 8–11 show the average confusion matrix for GTCCs derived from SVD, followed by Precision, Sensitivity, Specificity, and F1-score, as well as the average confusion matrix calculated for MFCCs derived from SVD, followed by Precision, Sensitivity, Specificity, and F1-score, respectively (Table 9, Table 10, Table 11).

Statistical significance

To assess the statistical significance of the obtained results, the *t*-test was performed. Although the significance level is typically set to 0.05, the significance level of 0.01 was chosen here to strengthen the hypothesis and minimise the chance of the occurrence of Type I error.

By applying a paired *t*-test, we calculated the mean difference between the accuracies of the two methods to be 21.824 and 13.572 for the custom dataset and SVD, respectively, with a standard deviation of approximately 1.21 and 2.71. The t-value calculated for the custom dataset was approximately 40.21, while the one for SVD – 11.18. With four degrees of freedom, compared against the critical t-value of 4.604 at the significance level of 1 %, we found the application of GTCCs as feature extraction method for classification of audio signals based on an underlying laryngeal pathology to be statistically significantly better than MFCCs. This analysis provides robust evidence supporting the superior performance of GTCC over MFCC in the classification task.

Discussion and conclusion

In this paper, we presented an accurate audio-based 3-class laryngeal pathology classification system based on Gammatone Cepstral Coefficients as the feature extraction method and CNNs as the classification model. The investigated laryngeal pathologies included cancerous and precancerous lesions on one hand and non-cancerous neuromuscular disorders on the other. The proposed system was tested on two different datasets; a custom dataset created for this research, as well as a widely used database in the field of vocal tract pathology classification – the Saarbruecken Voice Database (SVD). To address the gap in the literature related to the lack of reporting on the performance of continuous speech in laryngeal pathology classification systems, we used audio recordings of speech rather than sustained vowel phonation. Thus, the proposed approach differs fundamentally from prior work in that it focuses on speech-based, rather than phonation-based, feature extraction and employs GTCCs to better capture perceptually relevant information aligned with human auditory processing. This design choice was motivated by the hypothesis that the dynamic glottal behaviour and articulatory transitions present in speech carry additional diagnostic cues absent in sustained vowels.

While the proposed system demonstrates strong performance in distinguishing between precancerous and cancerous lesions, neuromuscular disorders, and healthy voice signals, it is intended as an assistive, pre-screening tool designed to complement existing clinical diagnosis instead of replacing it. The system is designed to provide an opportunity for a quicker critical illness flagging. In cases where the output of the automated system differs from a clinician's assessment, the clinician's judgment remains decisive, with the system serving to provide additional, objective acoustic evidence rather than a definitive diagnosis.

Compared to other studies reporting high classification accuracy with ERB spectrum-derived features [14,21,24], we found that using continuous speech signals instead of sustained phonation can improve the detection accuracy of precancerous and cancerous lesions. With GTSL features derived from sustained phonation, the average accuracy of detecting "structural" conditions in SVD, including some of precancerous and cancerous categories, was reported at 59.61 % [14]. In contrast, our system, utilising GTCCs from continuous speech signals, achieved an accuracy of 80.76 % on a custom dataset and 72.86 % on the SVD dataset. The observed improvement in classification accuracy across both datasets confirms the hypothesis that the natural variability

Table 7
Average precision (Pr), sensitivity (RC), specificity (Sp), and F1-score reported as % achieved by MFCCs derived from the custom dataset.

Parameter	Cancerous and Precancerous	Neuromuscular	Healthy
Precision	58.70 %±10.08	69.80 % ±16.85	67.46 %±9.87
Sensitivity	45.63 %±19.18	64.02 %±20.05	79.99 % ±15.49
F1-Score	48.75 %±13.53	64.77 %±11.80	71.69 %±5.52
Specificity	82.50 %±12.52	83.69 %±14.63	78.64 %±12.32

Table 6
Average confusion matrix calculated for the proposed system fed with MFCCs derived from the custom dataset.

MFCC on custom data		PREDICTED CLASS		
		Cancerous and Precancerous	Neuromuscular	Healthy
TRUE CLASS	Cancerous and Precancerous	**46.30 %**	27.97 %	25.72 %
	Neuromuscular	19.61 %	**63.99 %**	16.40 %
	Healthy	16.40 %	4.18 %	**79.42 %**

Table 8
Average confusion matrix calculated for the proposed system fed with GTCCs derived from the SVD.

GTCC on SVD		PREDICTED CLASS		
		Cancerous and Precancerous	Neuromuscular	Healthy
TRUE CLASS	Cancerous and Precancerous	**72.86 %**	15.00 %	12.14 %
	Neuromuscular	6.07 %	**66.79 %**	27.14 %
	Healthy	3.57 %	14.29 %	**82.14 %**

Table 9
Average precision (Pr), sensitivity (RC), specificity (Sp), and F1-score reported as % achieved by GTCCs derived from the SVD.

Parameter	Cancerous and Precancerous	Neuromuscular	Healthy
Precision	88.50 %±2.76	70.28 %±4.87	67.82 %±2.60
Sensitivity	72.86 %±5.84	66.79 %±8.05	82.14 %± 5.36
F1-Score	79.74 %±3.16	68.03 %±1.93	74.16 %±1.58
Specificity	95.18 %±1.49	85.36 %±5.81	80.36 %±3.16

of glottal and articulatory motion during speech introduces diagnostic characteristics not observable in prolonged vowel sounds.

The proposed classification method, when tested with the custom dataset, achieved the average accuracy of 85.04 % ±1.23 in classification of audio signals amongst three selected laryngeal pathologies. We compared the application of GTCCs against the more prevalently used MFCCs – the designed CNN classifier fed with MFCCs derived from the customs dataset achieved the average accuracy of 63.22 % ±1.62. When tested on SVD, the proposed system achieved an average accuracy of 73.93 % ±1.42 and 60.36 % ±2.44 using GTCCs and MFCCs, respectively. A *t*-test was used to demonstrate the statistical significance of the differences. for the purposes of audio-based laryngeal pathology classification.

As the key limitation of this study, we identified the imbalance present in the datasets. Despite testing the system on two separate datasets, both exhibited significant disparities in the number of participants across different categories. While addressed with a random oversampling method, this imbalance may have hindered the algorithm's ability to generalise effectively across all classes. However, the imbalance in the custom dataset was notably less severe than that observed in SVD, reflecting a more proportionate representation of pathological and control cases and, consequently, a more balanced evaluation of the proposed approach. Nevertheless, in both datasets, the sample size of each class was relatively low – particularly the pathological signals in SVD and the healthy in the custom dataset. Deep learning models typically require extensive and diverse data to achieve optimal performance. The limited size of the dataset could have affected the model's learning capacity and predictive accuracy.

It is also important to acknowledge that the metrics reported in this study were computed on one-second signal segments rather than on a per-subject basis. While segment-level analysis provides a robust means of increasing the number of training instances and assessing within-recording consistency, it may inflate the effective sample size. Applying a majority-voting scheme at test time – where predictions from all segments belonging to the same recording are aggregated to infer a subject-level diagnosis – was not feasible in the present study due to the limited number of available participants. Nevertheless, the high consistency observed among segment-level classifications supports the inference that whole-recording diagnoses would remain stable under such a scheme. Future work will therefore extend this methodology toward subject-level diagnostic prediction once a medically sufficient sample size is available. The intended future work will focus on acquiring a more balanced and comprehensive dataset to enhance the robustness and generalisability of the system. Therefore, future work will focus on investigating additional speech tasks and exploring multimodal approaches combining acoustic features with physiological data, such as bioimpedance measurements obtained with electroglottography [23].

The conclusion of this study is that a GTCC-fed CNN model based on audio recordings of human voice is an effective method for the classification of laryngeal pathologies. In particular, the results point strongly to the possibility of developing an automated pre-screening method for cancerous and pre-cancerous conditions. Further studies into more extensive medical datasets would be required towards this goal. In addition, investigations on including additional modalities (e.g. electroglottography signals) are under way, with the aim of further increasing the classification performance.

CRediT authorship contribution statement

Julia Zofia Tomaszewska: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Wojciech Kukwa:** Writing – review & editing, Supervision, Resources, Data curation. **Apostolos Georgakis:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Table 10
Average confusion matrix calculated for the proposed system fed with MFCCs derived from the SVD.

MFCC on SVD		PREDICTED CLASS		
		Cancerous and Precancerous	Neuromuscular	Healthy
TRUE CLASS	Cancerous and Precancerous	**60.81 %**	22.97 %	16.22 %
	Neuromuscular	31.42 %	**50.68 %**	17.91 %
	Healthy	14.86 %	15.88 %	**69.26 %**

Table 11
Average precision (Pr), sensitivity (RC), specificity (Sp), and F1-score reported as % achieved by MFCCs derived from the SVD.

Parameter	Cancerous and Precancerous	Neuromuscular	Healthy
Precision	56.87 %±2.51	58.40 %±11.58	69.34 %±9.94
Sensitivity	61.43 %±6.39	50.08 %±8.41	69.56 %±11.82
F1-Score	58.91 %±3.61	52.99 %±5.20	68.39 %±4.87
Specificity	76.65 %±3.24	80.62 %±8.86	83.27 %±8.63

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research has been conducted as part of a PhD at the University of West London under the Vice Chancellor Scholarship Scheme.

Data availability

Data will be made available on request.

References

[1] N. Roy, R.M. Merrill, S.D. Gray, E.M. Smith, Voice disorders in the general population: prevalence, risk factors, and occupational impact, Laryngoscope 115 (11) (2005) 1988–1995. Nov.

[2] L. Sulica, Laryngoscopy, stroboscopy and other tools for the evaluation of voice disorders, Otolaryngol. Clin. North Am. 46 (1) (2013) 21–30. Feb. 2013.

[3] J.I. Godino-Llorente, P. Gomez-Vilda, Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, IEEE Trans. Biomed. Eng. 51 (2) (2004) 380–384. Feb.

[4] R.J. Moran, R.B. Reilly, P. de Chazal, P.D. Lacy, Telephony-based voice pathology assessment using automated speech analysis, IEEE Trans. Biomed. Eng. 53 (3) (2006) 468–477. March.

[5] C. Peng, W. Chen, X. Zhu, B. Wan, D. Wei, Pathological voice classification based on a single vowel's acoustic features, in: 7th IEEE International Conference on Computer and Information Technology (CIT 2007), Japan, Aizu-Wakamatsu, 2007, pp. 1106–1110.

[6] P. Henriquez, J.B. Alonso, M.A. Ferrer, C.M. Travieso, J.I. Godino-Llorente, F. Diaz-de-Maria, Characterization of healthy and pathological voice through measures based on nonlinear dynamics, IEEE Trans. Audio Speech. Lang. Process. 17 (6) (2009) 1186–1195. Aug.

[7] J.D. Arias-Londoño, J.I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, G. Castellanos-Domínguez, Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients, IEEE Trans. Biomed. Eng. 58 (2) (2011) 370–379. Feb.

[8] M. Markaki, Y. Stylianou, Voice pathology detection and discrimination based on modulation spectral features, IEEE Trans. Audio Speech. Lang. Process. 19 (7) (2011) 1938–1948. Sept.

[9] D. Hemmerling, A. Skalski, J. Gajda, Voice data mining for laryngeal pathology assessment, Comput. Biol. Med. 69 (1) (2016) 270–276. Feb.

[10] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, M. Farahat, K. Malki, An investigation of Multidimensional Voice Program parameters in three different databases for voice pathology detection and classification, J. Voice 31 (1) (2017), 113.e9-113.e18Jan.

[11] P. Harar, J.B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, Z. Smekal, Voice pathology detection using deep learning: a preliminary study, in: 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), Funchal, Portugal, 2017, pp. 1–4.

[12] M. Borsky, D.D. Mehta, J.H. Van Stan, J. Gudnason, Modal and nonmodal voice quality classification using acoustic and electroglottographic features, IEEE ACM Trans. Audio Speech Lang. Process. 25 (12) (Dec. 2017) 2281–2291.

[13] S.S. Wang, C.T. Wang, C.C. Lai, Y. Tsao, S.H. Fang, Continuous speech for improved learning pathological voice disorders, IEEE Open. J. Eng. Med. Biol. 3 (2022) 25–33.

[14] C. Zhou, Y. Wu, Z. Fan, X. Zhang, D. Wu, Z. Tao, Gammatone spectral latitude features extraction for pathological voice detection and classification, Appl. Acoust. 185 (2022) 108417. Jan.

[15] H.C. Kuo, Y.P. Hsieh, H.H. Tseng, C.T. Wang, S.H. Fang, Y. Tsao, Toward real-world voice disorder classification, IEEE Trans. Biomed. Eng. 70 (10) (2023) 2922–2932. Oct.

[16] N.M.A.A. Latiff, F.T. Al-Dhief, N.F.S.M. Sazihan, M.M. Baki, N.N.N.A. Malik, M.A. A. Albadr, A.H. Abbas, Voice pathology detection using machine learning algorithms based on different voice databases, Results. Eng. 25 (2025) 103937. Mar. 2025.

[17] R. Islam, E. Abdel-Raheem, M. Tarique, 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), Ras Al Khaimah, United Arab Emirates, 2022, pp. 127–131.

[18] X. Valero, F. Alias, Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification, IEEE Trans. Multimed. 14 (6) (2012) 1684–1689. Dec.

[19] S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch, J. Acoust. Soc. Am. 8 (3) (1937) 185–190. Jan.

[20] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice, An efficient auditory filterbank based on the gammatone function, in: a meeting of the IOC Speech Group on Auditory Modelling at RSRE 2, 1987. Dec.

[21] D. Kumar, U. Satija, P. Kumar, Analysis and classification of electroglottography signals for the detection of speech disorders, in: 2023 National Conference on Communications (NCC), Guwahati, India, 2023, pp. 1–6.

[22] D. Bonet-Sola, R.M. Alsina-Pages, A comparative survey of feature extraction and machine learning methods in diverse acoustic environments, Sensors 21 (4) (2021) 1274. Feb.

[23] J.Z. Tomaszewska, A. Georgakis, Electroglottography in medical diagnostics of vocal tract pathologies: a systematic review, J. Voice (2023). Dec.

[24] R. Islam, E. Abdel-Raheem, M. Tarique, A novel pathological voice identification technique through simulated cochlear implant processing systems, Appl. Sci. 12 (5) (2022) 2398. Feb.

[25] J.Y. Lee, Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the Saarbruecken voice database, Appl. Sci. 11 (15) (2021) 7149. Aug.

[26] W.J. Barry, M. Putzer, Saarbrucken Voice Database. Institute of Phonetics, University of Saarland, Institute of Phonetics University of Saarland, 2007.

Dr Julia Zofia Tomaszewska, MSc, PhD, Lecturer in Audio Engineering at the University of West London. Her area of expertise includes digital signal processing, biomedical acoustics, digital audio engineering and machine learning systems. Currently involved in various projects concerning digital pre-screening methods, including the development of a multimodal classification system for laryngeal pathologies.

Asst. Prof. Wojciech Kukwa, MD, PhD, ENT specialist and the head of the Otorhinolaryngology Clinic of the Medical University of Warsaw in the Czerniakowski Hospital in Warsaw. His research focuses on sleep breathing disorders, including obstructive sleep apnoea in both adults and children. He co-founded the Polish Healthy Sleep Foundation and is involved with the Clebre startup, which develops audio-based diagnostic tools for sleep-disordered breathing. He has published over 100 papers in international and national journals.

Assoc. Prof. Apostolos Georgakis, MSc, PhD, An associate professor and a PhD supervisor at the University of West London. He joined the School of Computing and Engineering at UWL in 2017. Prior to that he was with the Division of Engineering at King's College London. His area of expertise includes time-frequency analysis with applications in biomedical signal processing, information systems and mathematics.