UNIVERSITY OF WEST LONDON
The Career University

UWL REPOSITORY

repository.uwl.ac.uk

Predicting households risks of waterborne diseases in flood-affected areas using machine learning: A study of Noakhali and Feni, Bangladesh

Miah, Md. Mamun, Hossain, Kabir, Khan, Hafiz T.A. ORCID logoORCID: https://orcid.org/0000-0002-1817-3730 and Ali, H.M. Shahadat (2026) Predicting households risks of waterborne diseases in flood-affected areas using machine learning: A study of Noakhali and Feni, Bangladesh. PLOS Water, 5 (8).

https://doi.org/10.1371/journal.pwat.0000542

This is the Published Version of the final output.

UWL repository link: https://repository.uwl.ac.uk/id/eprint/15581/

Alternative formats: If you require this document in an alternative format, please contact: open.research@uwl.ac.uk

Copyright: Creative Commons: Attribution 4.0

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Rights Retention Statement:

PLOS Water

RESEARCH ARTICLE

Predicting households risks of waterborne diseases in flood-affected areas using machine learning: A study of Noakhali and Feni, Bangladesh

Md. Mamun Miah[1,2*], **Kabir Hossain**[1], **Hafiz T. A. Khan**[3,4], **H. M. Shahadat Ali**[5]

1 Department of Statistics, Noakhali Science and Technology University, Noakhali, Bangladesh, **2** Bangladesh Institute for Research and Statistical Training, Dhaka, Bangladesh, **3** Professor of Public Health Statistics, University of West London (UWL), Ealing, Brentford, United Kingdom, **4** Professorial Fellow, Oxford Institute of Population Ageing, University of Oxford, United Kingdom, **5** Department of Applied Mathematics, Noakhali Science and Technology University, Noakhali-, Bangladesh

* mamun.stat@nstu.edu.bd

Check for updates

OPEN ACCESS

Citation: Miah MM, Hossain K, Khan HTA, Ali HMS (2026) Predicting households risks of waterborne diseases in flood-affected areas using machine learning: A study of Noakhali and Feni, Bangladesh. PLOS Water 5(8): e0000542. https://doi.org/10.1371/journal.pwat.0000542

Editor: Jiafu Li, University of South Carolina, UNITED STATES OF AMERICA

Received: March 6, 2026

Accepted: July 31, 2026

Published: August 28, 2026

Copyright: © 2026 Miah et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The dataset is publicly available at https://doi.org/10.6084/m9.figshare.33059051.

Funding: This research was supported by the Research Cell, Noakhali Science and Technology University, Noakhali, Bangladesh

Abstract

Waterborne diseases remain a major public health challenge, especially in low- and middle-income countries like Bangladesh. This study aimed to identify household level predictors of waterborne diseases and develop an optimized, explainable machine learning framework for predicting disease risk in flood-affected communities. This cross-sectional study was conducted in two flood-affected districts in Bangladesh, using face-to-face interviews with 501 respondents selected through simple random sampling to collect household-level data. Following data preprocessing, correlation analysis, variance inflation factor assessment, mutual information analysis, and Recursive Feature Elimination with Cross-Validation (RFECV) were performed for feature selection. Five machine learning algorithms—Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost) were optimized using randomized hyperparameter tuning with five-fold cross-validation. Model performance was evaluated on an independent test set using accuracy, precision, recall, specificity, F1-score, receiver operating characteristic area under the curve (ROC-AUC), bootstrap confidence intervals, and SHAP-based interpretation. The prevalence of waterborne disease in the flood affected area was found to be 63.7% (n = 319). RFECV retained ten important predictors: age, gender, monthly household income, household size, flood frequency, flood severity, source of drinking water, availability of medical assistance, water purification method, and community preparedness for preventing waterborne diseases. Among the optimized machine learning models, the Random Forest classifier showed the most balanced predictive performance on the independent test set, achieving an accuracy of 61.4%, precision of 71.9%, recall of 64.1%, specificity of 56.8%, F1-score of 67.8%, and ROC-AUC of 0.643. SHAP analysis identified flood

(to MMM). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

severity, community preparedness, water purification practices, and socio-economic factors as the most influential predictors. Although stacking and weighted soft voting ensemble models were additionally evaluated, neither outperformed the Random Forest. The proposed explainable framework may support targeted interventions and early warning strategies after external validation. All the analyses were conducted using Python.

1. Introduction

Waterborne diseases continue to be a major global public health concern, especially in low- and middle-income countries (LMICs), with limited access to clean drinking water, adequate sanitation, and good hygiene habits [1]. Frequent flooding, inadequate sanitation systems, and contamination of water sources all contribute significantly to the persistence and spread of waterborne pathogens [2]. The most prevalent waterborne illnesses are cholera, shigella, typhoid, poliomyelitis, hepatitis A and E, and diarrheal diseases [3]. More than 2 billion people lived in water-stressed nations in 2021; this problem is predicted to get worse due to population growth and climate change. By 2022, at least 1.7 billion people globally were consuming drinking water sources contaminated with feces, which posed a major health risk [4]. Waterborne diseases account for an estimated 3.5 million deaths annually, including approximately 2.2 million deaths among children, highlighting the disproportionate burden borne by vulnerable populations [5]. Climate change, rapid population growth, and increasing water scarcity are expected to further exacerbate these challenges, particularly in flood-prone regions.

Bangladesh has a higher prevalence of waterborne diseases than other nations because of its geographic, environmental, and socioeconomic characteristics. Diarrheal diseases remain a leading cause of morbidity and mortality among children under five years of age, with a mortality rate of 12.69 deaths per 100,000 children reported in 2019 [6]. Between 2017 and 2020, surveillance data documented 73,606 cases of cholera, 38,472 cases of typhoid fever, 2,510 cases of hepatitis A, and 1,643 cases of amoebiasis across Bangladesh [7]. This high burden is largely driven by inadequate water, sanitation, and hygiene (WASH) conditions, with 65.5 million (38.3%) lacking basic hygiene, 70 million (40.9%) relying on contaminated water, and 70 million (40.7%) without proper toilets conditions that allow floods to mix with waste and trigger widespread outbreaks [8]. Rural and coastal areas are disproportionately affected, as restricted healthcare access and delayed medical intervention increase morbidity and fatality rates during flood occurrences [9]. Understanding household-level risk factors in these vulnerable settings is therefore essential for developing effective prevention and mitigation strategies.

The occurrence of waterborne diseases is influenced by a complex interaction of environmental, socioeconomic, infrastructural, and behavioral factors. These include low levels of public awareness regarding hygienic practices, inadequate sanitation facilities, a lack of access to safe drinking water, and poor water quality

[10]. Socioeconomic factors, including living conditions, education level, and poverty, are also very important [11,12]. The transmission and spread of waterborne diseases are being greatly impacted by climate change, which is changing the quality and availability of water and raising health risks for all impacted populations. This is due to rising temperatures, changed rainfall patterns, droughts, flooding, and sea level rise [13]. Rapid and unplanned urbanization also contributes to overcrowding, inadequate sanitation infrastructure, and increased pressure on water supply systems, creating favorable conditions for waterborne disease outbreaks [14].

Recent advances in machine learning (ML) have created new opportunities for improving disease surveillance and public health decision-making by enabling accurate prediction of disease risks using complex, multidimensional datasets [15]. Unlike conventional statistical approaches that often require predefined relationships among variables, machine learning algorithms can capture nonlinear interactions and complex patterns among environmental, demographic, and behavioral factors [16,17]. Consequently, ML has been increasingly applied to infectious disease prediction, early warning systems, and public health resource allocation. However, the majority of earlier research in Bangladesh has concentrated on climate and environmental data aggregates rather than household-level forecasts, thus ignoring the variation in risk variables within various areas [7,18]. Household-level prediction models that integrate environmental, socioeconomic, infrastructural, and behavioral characteristics remain limited. Moreover, many existing studies rely on direct comparisons of machine learning algorithms without systematic feature selection, comprehensive hyperparameter optimization, model calibration, explainability analysis, or assessment of clinical utility, limiting both predictive robustness and interpretability.

Despite ongoing efforts by government and non-governmental organizations to improve water, sanitation, and hygiene (WASH) infrastructure, waterborne diseases remain highly prevalent in rural areas and densely populated urban slums [19]. Developing reliable household-level prediction models could facilitate early identification of vulnerable populations and support targeted public health interventions before disease outbreaks occur. Therefore, this study aimed to develop and evaluate an optimized and explainable machine learning framework for predicting household-level waterborne disease risk in two flood-prone districts of Bangladesh (Feni and Noakhali). The proposed framework integrates systematic feature selection using Recursive Feature Elimination with Cross-Validation (RFECV), comprehensive hyperparameter optimization, bootstrap validation, and SHAP-based model interpretation to identify important predictors and compare the predictive performance of five commonly used machine learning algorithms. By providing an interpretable and rigorously validated prediction framework, this study offers evidence that may support risk-based surveillance, targeted interventions, and informed public health decision-making in flood-affected communities.

2. Methodology

2.1 Ethics statement

Ethical approval for this study was obtained from the Noakhali Science and Technology University Ethical Committee (NSTUEC) (Reference No. NSTU/SCI/EC/2024/268). The study was conducted in accordance with the principles of the Declaration of Helsinki and its subsequent amendments. Written informed consent was obtained from all participants before data collection. Participants were informed about the objectives of the study, the voluntary nature of their participation, the confidentiality of their responses, and their right to withdraw from the study at any time without any consequences.

2.2. Study area

Bangladesh (23.6850° N - 90.3563° E) is the world's largest deltaic south Asian country, located in the tropical monsoon region in the northeastern part of the Indian subcontinent. This study was carried out in two flood affected areas (Noakhali and Feni) in Bangladesh.

2.3. Study design and participants

This study employed a community-based cross-sectional design to identify household level predictors of waterborne diseases and develop machine learning (ML) models for predicting household level disease risk in flood affected communities in Bangladesh. The study was conducted in selected communities in the districts of Noakhali and Feni, where concerns regarding drinking water contamination and waterborne disease transmission increased following a flash flood event that occurred between 20 and 22 August 2024. The flooding affected local water sources and sanitation systems, creating potential risks for waterborne diseases transmission in the affected communities.

The overall conceptual framework and analytical workflow of the study are presented in Fig 1 and Fig 2, respectively.

Participants were prospectively recruited between 5 October 2024 and 30 December 2024 using a structured questionnaire administered through face-to-face interviews. Eligible participants were adults aged 18 years or older who were permanent residents of the selected communities and provided written informed consent. Individuals who were unable to provide informed consent or complete the questionnaire were excluded from the study.

A total of 501 respondents participated in the survey. Following data screening and quality assurance procedures, including checks for completeness, consistency, and duplicate responses, all 501 observations were retained for the final analyses. The final dataset contained 319 households reporting at least one waterborne disease event and 182 households without reported disease, providing an adequate number of events for prediction model development.

2.4 Data collection

Data were collected through face-to-face interviews using a structured questionnaire administered by trained data collectors. Before the survey, the questionnaire was pretested in a community with characteristics similar to those of the study

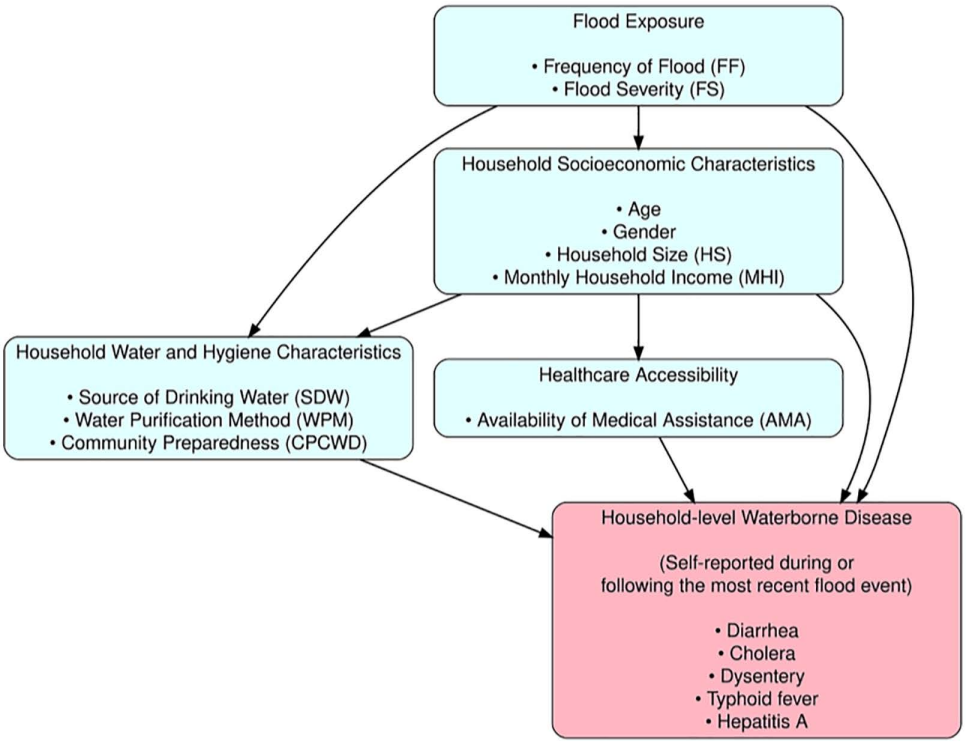


Fig 1. Overall conceptual framework of the study.

https://doi.org/10.1371/journal.pwat.0000542.g001

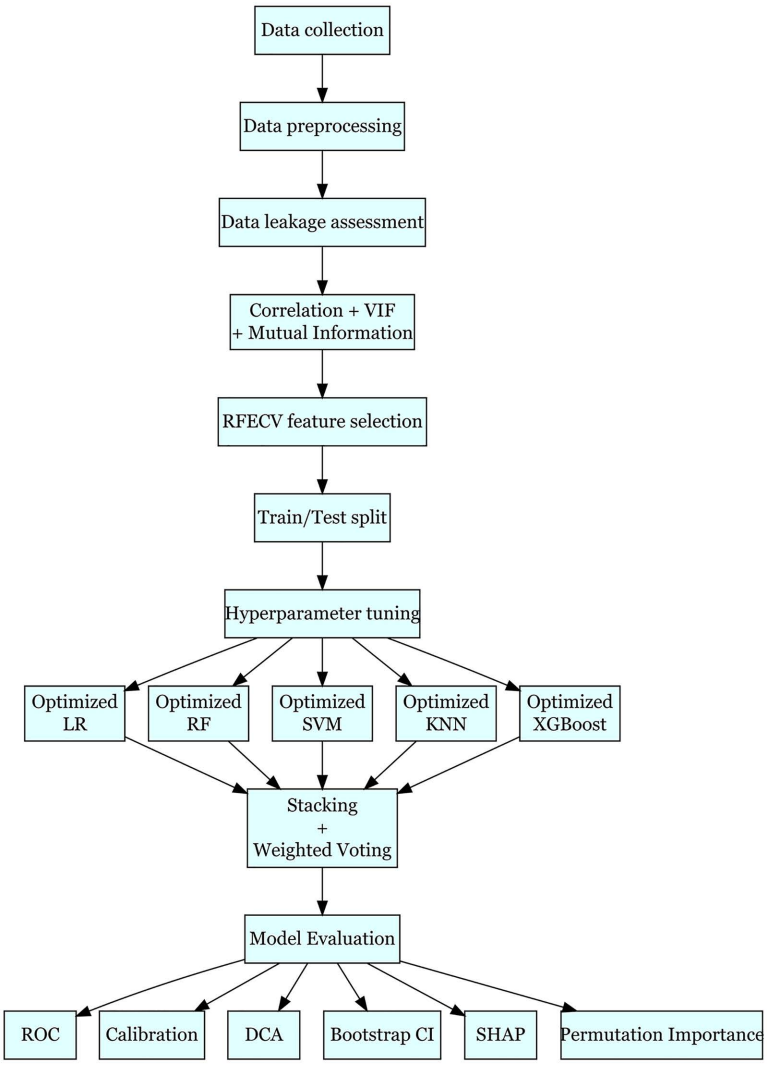


Fig 2. Analytical workflow of the study.

https://doi.org/10.1371/journal.pwat.0000542.g002

population to assess its clarity, relevance, and comprehensibility. Necessary modifications were made before the final data collection.

The questionnaire consisted of five sections covering: (i) demographic characteristics, (ii) household socioeconomic information, (iii) flood exposure and environmental conditions, (iv) drinking water sources, water purification practices, healthcare accessibility, and community preparedness, and (v) household history of waterborne diseases.

Participants were asked whether any household member had experienced diarrhea, cholera, dysentery, typhoid fever, or hepatitis A during or following the most recent flood event. Disease information was obtained through participants' self-reports using the structured questionnaire and was not verified through clinical examination, laboratory diagnosis, or medical records. An additional response option ("Others") was included to capture other self-reported water-related illnesses; however, no participants selected this option. Consequently, the outcome variable was based exclusively on the five predefined waterborne diseases.

To ensure data quality, interviewers received standardized training on the study objectives, questionnaire administration, ethical considerations, and interview procedures. Completed questionnaires were reviewed daily for completeness and internal consistency before data entry.

2.5 Variables and operational definitions

The study considered demographic, socioeconomic, environmental, healthcare accessibility, and behavioral characteristics as candidate predictors of household-level waterborne disease. Table 1 presents the operational definitions, measurement scales, and coding schemes for all variables included in the analysis.

Source of Drinking Water (SDW) and Water Purification Method (WPM) were converted into binary thematic indicators to reduce category sparsity, improve model stability, and facilitate interpretation while preserving their public health relevance. Household-level waterborne disease (SWD) was defined as the self-reported occurrence of at least one predefined waterborne disease during or following the most recent flood event and was coded as a binary outcome (0 = No, 1 = Yes).

2.6 Data preprocessing and cleaning

Prior to model development, the dataset was examined for completeness, consistency, duplicate observations, and coding errors. A total of 501 household survey responses were collected from flood-affected communities in Noakhali and Feni districts. Following data screening, all 501 records were retained for the final analyses because no records met the exclusion criteria for duplication, logical inconsistency, or incomplete questionnaires.

Table 1. Description of study variables.

Type	Abbreviation	Description	Levels	Coding
Environmental	FF	Frequency of flood	Ordinal	0 = Rarely, 1 = Once a year, 2 = Twice a year, 3 = More than twice a year
	FS	Flood severity	Ordinal	0 = Not affected, 1 = Mildly affected, 2 = Severely affected
	SDW	Source of drinking water	Thematic indicator	1 = Any drinking water source used, 0 = Otherwise
Socio-economic	Gender	Gender of the respondent	Male, Female	1 = male, 0 = Female
	Age	Age of the respondent	Continuous (in years)	
	MHI	Monthly household income	Category	1 = < 10,000 BDT, 2 = 10,000–20,000 BDT, 3 = > 20,000 BDT
	HS	Household size	Category	0 = 1–3, 1 = 4–6, 2 = More than 6
Infrastructural	MAWD	Most affected household members	Category	0 = Children (<12), 1 = Adults (12–60), 2 = Elderly (>60)
	AMA	Availability of medical assistance	Binary category	0 = No, 1 = Yes
	MA	Seek medical assistance	Binary category	0 = No, 1 = Yes
	TT	Type of treatment received	Category	0 = Hospital, 1 = Clinic, 2 = Self-medication, 3 = Traditional healer
	HRAI	Hygiene-related awareness or intervention	Binary category	0 = No, 1 = Yes
Health outcome	SWD	Suffered from waterborne disease	Binary category	0 = No, 1 = Yes
Behavioural and protective practices	WPM	Water purification method	Thematic indicator	1 = Any purification method used, 0 = Otherwise
	ACWD	Awareness about waterborne diseases	Binary category	0 = No, 1 = Yes
	CPCWD	Community Preparedness to prevent waterborne diseases	Binary category	0 = No, 1 = Yes

https://doi.org/10.1371/journal.pwat.0000542.t001

All variables were encoded according to the predefined coding scheme presented in Table 1. Age was retained as a continuous variable, whereas all remaining predictors were represented as binary, ordinal, or categorical variables using numerical encoding. The outcome variable, Suffered from Waterborne Disease (SWD), was defined as a binary classification variable (0 = No, 1 = Yes).

The dataset was randomly partitioned into training (80%) and testing (20%) subsets using stratified sampling to preserve the original class distribution of the outcome variable in both datasets. The training dataset was used for feature selection, hyperparameter optimization, and model development, whereas the independent testing dataset was reserved exclusively for final model evaluation. This separation provided an unbiased assessment of predictive performance and minimized optimistic performance estimates.

Because several machine learning algorithms, particularly Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), are sensitive to differences in predictor scales, continuous variables were normalized using Min–Max scaling. Scaling parameters were estimated exclusively from the training dataset and subsequently applied to the testing dataset to prevent information leakage. Tree-based algorithms, including Random Forest and Extreme Gradient Boosting (XGBoost), were trained using the same preprocessed datasets to ensure methodological consistency across all models.

Potential temporal data leakage was also considered during model development. Variables representing healthcare-seeking behaviors, such as seeking medical assistance (MA) and type of treatment received (TT), may occur after disease onset and therefore have the potential to encode post-outcome information. These variables were initially included as candidate predictors but were carefully evaluated during the feature selection process and were not retained in the final prediction models, thereby reducing the likelihood of temporal information leakage.

To assess redundancy among predictors, pairwise associations were examined using Spearman's rank correlation coefficients, and multicollinearity was evaluated using the Variance Inflation Factor (VIF). Although two variables (SDW and MHI) exhibited moderate to high VIF values, these variables were retained because the primary objective of the study was prediction rather than causal inference, and multicollinearity generally has a limited impact on the predictive performance of most machine learning algorithms, particularly tree-based methods. Furthermore, Recursive Feature Elimination with Cross-Validation (RFECV) was subsequently applied to identify the most informative subset of predictors while reducing redundancy and the potential for overfitting. In addition, Mutual Information (MI) analysis was performed to quantify the dependency between each predictor and the outcome variable, providing an initial assessment of feature relevance before feature selection. These preprocessing procedures improved data quality, reduced predictor redundancy, minimized the risk of information leakage, and established a robust foundation for subsequent machine learning analyses. Pairwise associations among predictors were visualized using a Spearman correlation heatmap (S1 Fig), while multicollinearity was assessed using Variance Inflation Factor (VIF) values (S3 Table).

2.7 Feature selection

To improve model robustness, reduce redundant predictors, and minimize the risk of overfitting, a systematic feature selection procedure was performed before model development. Variables with the potential to introduce data leakage—namely seeking medical assistance (MA), type of treatment received (TT), and most affected household member (MAWD)—were excluded because they represent post-disease events that would not be available at the time of prediction.

The remaining predictors were evaluated using three complementary feature selection methods: Mutual Information (MI), Recursive Feature Elimination with Cross-Validation (RFECV) with Logistic Regression (RFECV-LR), and RFECV with Random Forest (RFECV-RF). Mutual Information was used to quantify the dependency between each predictor and the outcome without assuming linear relationships. RFECV, implemented with five-fold stratified cross-validation, iteratively eliminated less informative variables to identify the optimal subset of predictors based on the highest mean area

under the receiver operating characteristic curve (AUC). Logistic Regression was employed to capture linear predictive relationships, whereas Random Forest was used to identify nonlinear relationships and interactions among predictors.

A consensus-based feature selection strategy was then applied. Predictors selected by at least two of the three methods were automatically retained, whereas predictors selected by only one method were further evaluated based on their epidemiological relevance, prior evidence, and contribution to predictive performance. As shown in Table 1, four predictors—Age, Household Size (HS), Flood Severity (FS), and Frequency of Flood (FF)—were selected by at least two methods and were therefore retained automatically. Six additional predictors—Gender, Monthly Household Income (MHI), Source of Drinking Water (SDW), Water Purification Method (WPM), Availability of Medical Assistance (AMA), and Community Preparedness to Prevent Waterborne Diseases (CPCWD)—were selected only by MI but were retained because of their established epidemiological importance and their contribution to model performance during evaluation. In contrast, Hygiene-related awareness or intervention (HRAI) and Aware of Causes of Waterborne Diseases (ACWD) were excluded because they were not selected by any feature selection method and did not improve predictive performance.

Consequently, ten predictors (Age, Gender, MHI, HS, FF, FS, SDW, WPM, AMA, and CPCWD) were included in the final model and were subsequently used to train and evaluate all machine learning models.

2.8 Hyperparameter optimization

To ensure a fair comparison among machine learning algorithms and improve predictive performance, hyperparameter optimization was performed independently for each classification model using RandomizedSearchCV with five-fold stratified cross-validation on the training dataset. Stratified cross-validation preserved the class distribution across folds and prevented information leakage from the independent testing dataset. Candidate hyperparameter ranges were defined according to published recommendations and the respective software documentation. For each algorithm, the optimal hyperparameter combination was selected based on the highest mean cross-validated ROC–AUC. Logistic Regression, Random Forest, Support Vector Machine, and Extreme Gradient Boosting were each evaluated using 50 randomly sampled hyperparameter combinations, whereas 30 combinations were evaluated for K-Nearest Neighbors because of its smaller search space. The optimized hyperparameter values and corresponding cross-validation performance are presented in S1 Table and S2 Table, respectively.

2.9 Machine learning model development

Five supervised machine learning algorithms were developed to predict household-level waterborne disease occurrence: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). These algorithms were selected because they represent diverse statistical and machine learning approaches capable of modeling both linear and complex nonlinear relationships between predictors and the outcome. Logistic Regression was included as a baseline statistical classifier because of its interpretability and widespread application in binary classification problems [20]. Random Forest, an ensemble tree-based algorithm, was selected for its ability to capture nonlinear relationships and complex interactions while reducing overfitting through bootstrap aggregation [21]. Support Vector Machine was employed because of its effectiveness in high-dimensional classification problems and its ability to identify optimal decision boundaries using kernel functions [22]. K-Nearest Neighbors was included as a distance-based learning algorithm that classifies observations according to the majority class among their nearest neighbors [23]. Extreme Gradient Boosting (XGBoost), a gradient boosting ensemble algorithm, was selected because of its high predictive accuracy, computational efficiency, and built-in regularization mechanisms for reducing overfitting [24].

Following data preprocessing and feature selection using Recursive Feature Elimination with Cross-Validation (RFECV), each model was trained using the optimized hyperparameters identified through the RandomizedSearchCV procedure described in Section 2.7. Model development was performed exclusively on the training dataset, while the

independent testing dataset was reserved for final model evaluation. To ensure a fair comparison, all algorithms were developed using the same training and testing partitions and the identical subset of predictors selected by RFECV.

In addition to the individual machine learning models, two ensemble learning approaches were evaluated to investigate whether combining multiple classifiers could improve predictive performance. A stacking ensemble was developed by integrating the predictions of the optimized base learners using a meta-classifier. Furthermore, a weighted soft voting ensemble combining the optimized Random Forest, Logistic Regression, and XGBoost models was constructed. Candidate weight combinations were evaluated using five-fold cross-validation on the training dataset, and the combination yielding the highest mean cross-validated area under the receiver operating characteristic curve (ROC–AUC) was selected. The optimal weight combination (4:4:2) achieved the highest cross-validated ROC–AUC (0.662) and was subsequently evaluated on the independent testing dataset.

2.10 Model evaluation

The predictive performance of each machine learning model was evaluated using the independent testing dataset, which was not used during model training or hyperparameter optimization. Model discrimination was primarily assessed using the Area Under the Receiver Operating Characteristic Curve (ROC–AUC), which quantifies the ability of a classifier to distinguish between households with and without waterborne disease across all possible classification thresholds [25,26]. ROC–AUC is independent of any single decision threshold and is particularly informative for evaluating binary classification models when class imbalance may influence conventional performance measures such as accuracy [26]. In addition, accuracy, precision, recall (sensitivity), specificity, and F1-score were calculated from the confusion matrix to provide a comprehensive evaluation of classification performance. Confusion matrices were generated for each model to summarize the numbers of true positives, true negatives, false positives, and false negatives.

Receiver operating characteristic (ROC) curves were constructed to compare the discriminative ability of all models. To assess the uncertainty of model performance, 95% confidence intervals for ROC–AUC were estimated using bootstrap resampling with 2,000 iterations. Pairwise comparisons of ROC–AUC values between models were further performed using DeLong's test to determine whether observed differences in discrimination were statistically significant (S4 Table).

Model calibration was evaluated using calibration curves, which compared predicted probabilities with observed event frequencies to assess the agreement between predicted and actual risks. The predictive performance of the optimized stacking ensemble and weighted soft voting ensemble models was evaluated using the same performance metrics to facilitate direct comparison with the individual machine learning algorithms [25].

For the best-performing Random Forest model, additional model-specific diagnostic analyses were conducted. The Out-of-Bag (OOB) score was used as an internal estimate of predictive performance based on bootstrap samples generated during Random Forest training. The OOB error curve was examined to evaluate model convergence and determine whether prediction error stabilized as the number of trees increased. Furthermore, feature importance based on the mean decrease in impurity (Gini importance) was calculated to provide a model-specific assessment of predictor importance and to complement the subsequent model interpretation analyses.

2.11 Model interpretation

To improve the interpretability of the machine learning models, the best-performing model (Random Forest) was further analyzed using Shapley Additive Explanations (SHAP). SHAP provides a unified game-theoretic framework for quantifying the contribution of each predictor to individual model predictions while preserving both the magnitude and direction of feature effects. A SHAP summary plot was generated to visualize the overall importance and influence of each predictor across all observations, whereas SHAP dependence plots were used to examine the relationship between the most influential predictor and the predicted probability of household-level waterborne disease.

To complement the SHAP analysis, permutation feature importance was calculated as a model-agnostic measure of predictor relevance. Feature importance was quantified by measuring the decrease in predictive performance after randomly permuting the values of each predictor while keeping all other variables unchanged. Predictors producing larger reductions in model performance were considered to have greater predictive importance. The combined use of SHAP and permutation feature importance provided a comprehensive interpretation of the optimized Random Forest model while avoiding causal interpretation of predictor effects.

2.12 Internal validation

The robustness and generalizability of the developed machine learning models were assessed through internal validation procedures. Initially, the dataset was divided into independent training (80%) and testing (20%) subsets using stratified sampling to preserve the class distribution of the outcome variable. Hyperparameter optimization and model development were performed exclusively on the training dataset using five-fold stratified cross-validation, thereby minimizing overfitting and improving model stability. The optimized models were subsequently evaluated on the independent testing dataset to obtain an unbiased estimate of predictive performance.

To further assess the stability and uncertainty of model discrimination, bootstrap resampling with 2,000 iterations was performed to estimate 95% confidence intervals (CIs) for the area under the receiver operating characteristic curve (ROC–AUC). Bootstrap samples were generated from the testing dataset with replacement, and the empirical percentile method was used to derive confidence intervals for each model. In addition, pairwise comparisons of ROC–AUC values were conducted using DeLong's test to determine whether differences in predictive discrimination between models were statistically significant.

Because data from an independent external population were not available, external validation was not performed. Therefore, the proposed models should be considered internally validated prediction models, and their generalizability to other geographic regions and populations should be confirmed through future external validation studies.

3. Result

3.1. Descriptive statistics

The mean age of the respondents was 42.78 years (SD = 15.26) (Table 2). Among them, 63.7% of participants had waterborne disease, while 36.3% did not have any type of waterborne disease (Fig 3). Approximately, 64.7% participants were male and 35.3% were female. The highest rates of waterborne disease were found in children (<12 years) (41.9%), followed by adults (12–60 years) (44.5%), and older (>60 years) (13.6%). More than one-half (53.8%) sought medical assistance, with the most common forms being hospital-based treatment (42.2%) and self-medication (39.8%). More than six (>6) people lived in the majority of households (65.3%). In the context of income, 20.8% earned < 10,000 BDT per month, whereas 40.8% earned > 20,000 BDT. More than half (50.5%) of the respondents said they experienced flooding once a year, while 49.4% reported being severely affected by the most recent flood. Approximately, 38.3% of respondents experienced flooding for more than two weeks, and 39.7% lacked access to safe drinking water. Furthermore, 48.9% stated that their sanitation systems were completely damaged, followed by partially damaged (38.5%) and not affected (12.6%). Approximately 61.7% lacked access to medical care. The majority (65.7%) reported no hygiene-related awareness or

Table 2. Descriptive statistics of independent variable (continuous).

Variable	Mean	SD	Min.	25%	50%	75%	Max.
Age	42.77844	15.26319	1	32	40	55	95

https://doi.org/10.1371/journal.pwat.0000542.t002

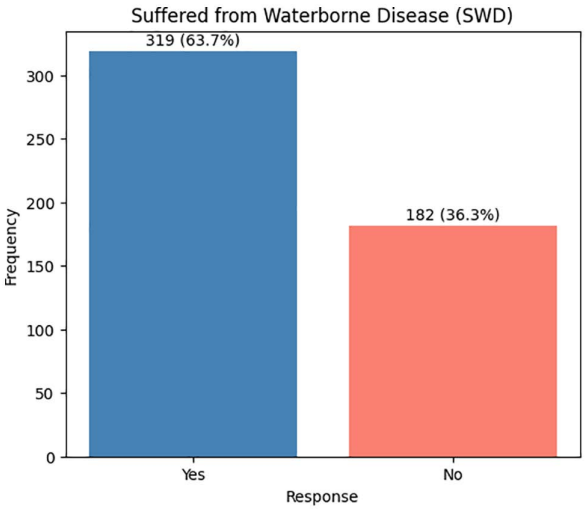


Fig 3. Frequency distribution of waterborne disease.

https://doi.org/10.1371/journal.pwat.0000542.g003

intervention. Furthermore, only 38.3% of the respondents had access to medical assistance, whereas 61.7% of them reported that it was not readily available or accessible (Table 3).

3.2 Chi-square analysis

The chi-square test revealed a significant correlation between several variables and SWD, with p-values < 0.05. The variables medical assistance sought (MA), most affected household members (MAWD), type of treatment received (TT), household size (HS), monthly household income (MHI), frequency of floods (FF), severity of flood impact (FS), and Awareness about waterborne diseases (ACWD) showed association with waterborne disease (Table 3).

3.3 Feature selection and Hyperparameter optimization

3.3.1 Feature selection using Mutual Information and Recursive Feature Elimination with Cross-Validation (RFECV). Feature selection was performed using a two-stage procedure comprising Mutual Information (MI) analysis followed by Recursive Feature Elimination with Cross-Validation (RFECV). Initially, MI was calculated to quantify the dependency between each predictor and the outcome variable and to provide an initial assessment of feature relevance. Subsequently, RFECV was applied using Logistic Regression and Random Forest as the base estimators with five-fold stratified cross-validation to identify the optimal subset of predictors while minimizing redundancy and reducing the risk of overfitting.

Both RFECV procedures consistently identified 10 predictors as the optimal subset for subsequent model development. The retained variables included Community Preparedness to Prevent Waterborne Diseases (CPCWD), Water Purification Method (WPM), Flood Severity (FS), Household Size (HS), Source of Drinking Water (SDW), Frequency of Flood (FF), Monthly Household Income (MHI), Gender, Availability of Medical Assistance (AMA), and Age. Variables not retained by RFECV demonstrated comparatively lower predictive contribution and were excluded from subsequent model development. The final selected predictors are presented in Table 4.

3.3.2 Hyperparameter optimization. Following feature selection, hyperparameter optimization was performed for Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, and XGBoost using RandomizedSearchCV with five-fold stratified cross-validation on the training dataset. Algorithm-specific search spaces

Table 3. Descriptive statistics and chi-square test results of some selected variables.

Variable	Frequency	Percentage	P value
Gender			0.8765
Male	324	64.7	
Female	177	35.3	
Household size (HS)			0.0005
1–3	63	12.6	
4–6	111	22.1	
>6	327	65.3	
Monthly household income (MHI)			0.0041
<10,000 BDT	104	20.8	
10,000–20,000 BDT	192	38.4	
>20,000 BDT	205	40.8	
Frequency of flood experiences (FF)			0.0018
Once a year	255	50.5	
Twice a year	177	35.1	
More than twice a year	68	13.5	
Rarely	1	0.2	
Severity of flood (FS)			<0.0001
Not affected	25	5	
Mildly affected	227	45	
Severely affected	249	49.4	
Awareness about waterborne disease (ACWD)			0.0048
Yes	299	59.68	
No	202	40.32	
Most affected household members (MAWD)			<0.0001
Children (<12 years)	210	41.9	
Adults (12–60 years)	223	44.5	
Elderly (>60 years)	68	13.6	
Seek medical assistance (MA)			<0.0001
Yes	269	53.8	
No	232	46.2	
Type of treatment received (TT)			<0.0001
Hospital-based treatment	212	42.2	
Self-medication	200	39.8	
Local health center/clinic	74	14.7	
Traditional healer	15	3	
Water purify method (WPM)			0.844
Any recommended method	274	54.69	
Others	227	45.31	
Awareness about waterborne diseases (ACWD)			0.05
Yes	299	59.7	
No	202	40.3	
Community Preparedness to prevent waterborne Diseases (CPCWD)			0.5146
Yes	163	32.5	
No	338	67.5	

https://doi.org/10.1371/journal.pwat.0000542.t003

Table 4. Mutual Information scores and variables retained following Recursive Feature Elimination with Cross-Validation (RFECV).

Feature	MI score	RFECV-LR	RFECV-RF	Final selection	Rationale
CPCWD	0.047	×	×	Yes	Epidemiological importance; improved model performance
WPM	0.040	×	×	Yes	Epidemiological importance
FS	0.034	√	√	Yes	Selected by all methods
HS	0.019	√	√	Yes	Selected by all methods
SDW	0.016	×	×	Yes	Improved model performance
FF	0.014	√	√	Yes	Selected by all methods
MHI	0.013	×	×	Yes	Epidemiological importance
Gender	0.007	×	×	Yes	Included as demographic covariate
AMA	0.004	×	×	Yes	Epidemiological importance
Age	0.000	√	√	Yes	Selected by RFECV methods
HRAI	0.000	×	×	No	No support from any method
ACWD	0.000	×	×	No	No support from any method

https://doi.org/10.1371/journal.pwat.0000542.t004

were predefined for each classifier, and the optimization objective was to maximize the mean cross-validated ROC–AUC while minimizing overfitting.

The optimized hyperparameter combinations identified through the search procedure were subsequently used to train the final prediction models. Among the individual classifiers, the optimized Random Forest model achieved the highest cross-validated ROC–AUC. The complete hyperparameter search space, optimal parameter values, and corresponding cross-validation results are presented in S1 Table and S2 Table, respectively.

3.4 Performance of machine learning models

The classification models were assessed using the test dataset, with confusion matrices (Fig 7), ROC curves (Fig 4) and corresponding evaluation metrics (Table 11) providing performance summaries.

3.4.1 Logistic regression. The optimized Logistic Regression (LR) model demonstrated satisfactory predictive performance for classifying household-level waterborne disease risk in the independent testing dataset. Following hyperparameter optimization using RandomizedSearchCV with five-fold cross-validation, the model achieved balanced discrimination and classification accuracy. The confusion matrix indicated that the model correctly classified a substantial proportion of households with and without waterborne disease. Although Logistic Regression provided a robust baseline model, its overall predictive performance was slightly lower than that of the optimized Random Forest model across most evaluation metrics. Detailed performance metrics are presented in Table 5.

3.4.2 Random forest. Among the individual machine learning algorithms, the optimized Random Forest (RF) model demonstrated the highest overall predictive performance. The model achieved the greatest accuracy, recall, F1-score, and ROC–AUC while maintaining an appropriate balance between sensitivity and specificity. Bootstrap-estimated 95% confidence intervals indicated stable predictive performance, and the ROC curve demonstrated superior discrimination compared with the remaining classifiers. These findings suggest that Random Forest was the most effective individual algorithm for predicting household-level waterborne disease risk within the present dataset. The detailed performance measures are summarized in Table 6.

3.4.3 Support Vector Machine (SVM). The optimized Support Vector Machine (SVM) classifier showed good predictive capability after hyperparameter optimization. The model produced satisfactory discrimination and balanced classification performance, although its accuracy and ROC–AUC were slightly lower than those of the Random Forest model. The SVM classifier nevertheless demonstrated competitive predictive performance and remained a suitable alternative for household-level waterborne disease classification. Detailed evaluation metrics are presented in Table 7.

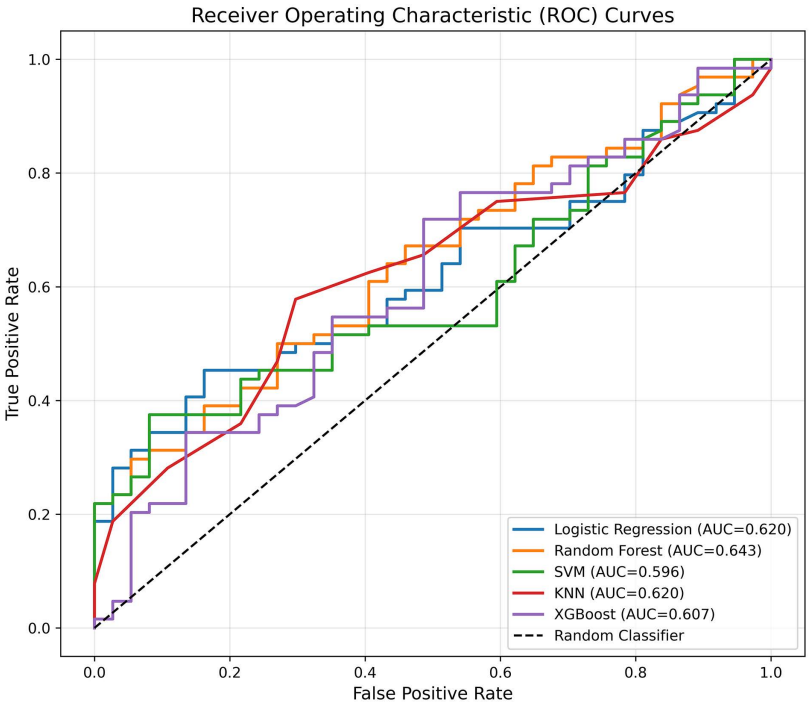


Fig 4. ROC curves for the applied classification models.

https://doi.org/10.1371/journal.pwat.0000542.g004

Table 5. Performance metrics of Logistic regression.

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC	Balance
Logistic Regression	0.554	0.694	0.531	0.595	0.602	0.620	Good specificity, weaker sensitivity

https://doi.org/10.1371/journal.pwat.0000542.t005

Table 6. Performance metrics of Random forest.

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC	Balance
Random Forest	0.614	0.719	0.641	0.568	0.678	0.643	Best balance

https://doi.org/10.1371/journal.pwat.0000542.t006

Table 7. Performance metrics of support vector machine.

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC	Balance
SVM	0.624	0.641	0.922	0.108	0.756	0.596	Poorly balanced

https://doi.org/10.1371/journal.pwat.0000542.t007

3.4.4 K-Nearest Neighbors (KNN). The optimized K-Nearest Neighbors (KNN) classifier demonstrated acceptable predictive performance in identifying households at risk of waterborne disease. Although hyperparameter optimization improved the model's performance, its classification accuracy and discrimination remained lower than those of the Random Forest and Logistic Regression models. The model exhibited reasonable sensitivity and specificity, indicating moderate predictive capability. Detailed performance results are shown in Table 8.

3.4.5 XGBoost. The optimized Extreme Gradient Boosting (XGBoost) model also demonstrated satisfactory classification performance following hyperparameter optimization. The model achieved competitive accuracy and

Table 8. Performance metrics of K-nearest neighbors.

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC	Balance
KNN	0.564	0.628	0.766	0.216	0.690	0.620	More positive-biased

https://doi.org/10.1371/journal.pwat.0000542.t008

ROC–AUC values but did not outperform the Random Forest classifier. Overall, XGBoost provided reliable predictive performance while exhibiting similar classification characteristics to the other optimized machine learning algorithms. The corresponding performance metrics are presented in Table 9.

3.4.6 Comparative model performance. A comparison of the optimized machine learning algorithms is presented in Table 10. Overall, the Random Forest model demonstrated the most balanced predictive performance, achieving the highest values for accuracy, recall, F1-score, and ROC–AUC among the individual classifiers. Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and XGBoost also demonstrated satisfactory predictive performance, although the differences in ROC–AUC between several models were relatively modest. Bootstrap-derived 95% confidence intervals indicated stable model performance, and pairwise comparisons using DeLong's test were performed to evaluate statistical differences in discrimination.

Receiver operating characteristic (ROC) curve analysis showed that the Random Forest model consistently achieved the highest discriminative ability across classification thresholds (Fig 4). Calibration curves indicated good agreement between predicted probabilities and observed outcomes for the optimized models (Fig 5). These findings support the selection of Random Forest as the best-performing individual classifier for subsequent interpretation and diagnostic analyses within the present dataset, while acknowledging that differences between models were generally modest and require external validation before broader implementation.

3.5 Ensemble learning performance

To investigate whether combining multiple classifiers could further improve predictive performance, two ensemble learning approaches, stacking and weighted soft voting, were developed using the optimized base learners. The ensemble models incorporated the five optimized machine learning algorithms (Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, and XGBoost) as base classifiers. Model development was performed using the training dataset, and predictive performance was evaluated on the independent testing dataset using the same evaluation metrics applied to the individual models.

Table 9. Performance metrics of XGBoost.

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC	Balance
XGBoost	0.604	0.658	0.781	0.297	0.714	0.607	More positive-biased

https://doi.org/10.1371/journal.pwat.0000542.t009

Table 10. Comparative performance metrics of all models.

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC	Balance
Random Forest	**0.614**	**0.719**	0.641	**0.568**	0.678	**0.643**	Best balance
Logistic Regression	0.554	0.694	0.531	**0.595**	0.602	0.620	Good specificity, weaker sensitivity
XGBoost	0.604	0.658	0.781	0.297	0.714	0.607	More positive-biased
KNN	0.564	0.628	0.766	0.216	0.690	0.620	More positive-biased
SVM	0.624	0.641	0.922	0.108	0.756	0.596	Poorly balanced

https://doi.org/10.1371/journal.pwat.0000542.t010

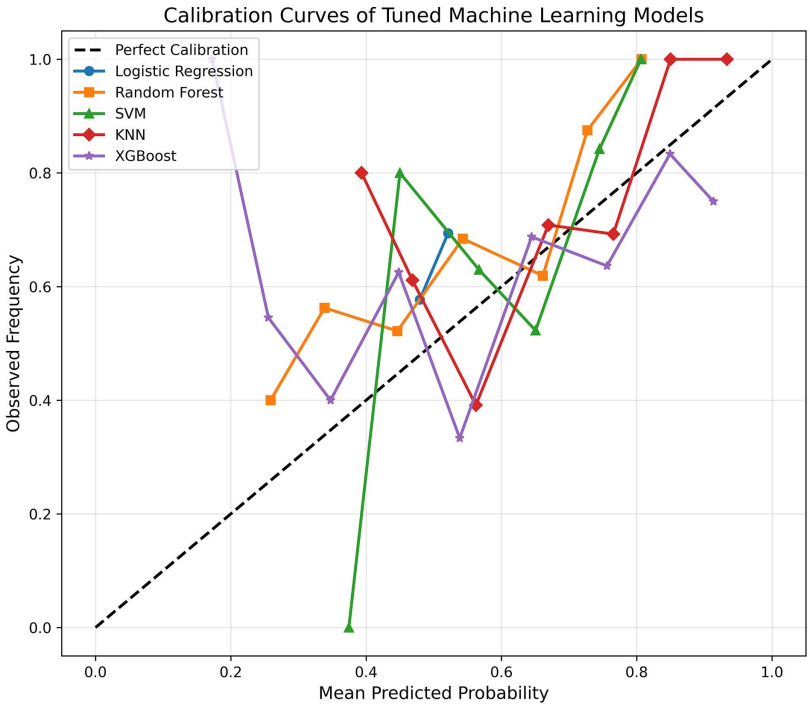


Fig 5. Calibration curve for all tuned ML models.

https://doi.org/10.1371/journal.pwat.0000542.g005

The stacking ensemble combined the predictions of the optimized base learners using a meta-classifier trained to integrate their outputs. This approach produced competitive classification performance and demonstrated good discrimination for predicting household-level waterborne disease risk. However, its overall predictive performance did not consistently exceed that of the optimized Random Forest model across all evaluation metrics.

Similarly, the weighted soft voting ensemble combined the predicted class probabilities of the optimized base classifiers using weights determined through cross-validation. The optimized weighting scheme improved the balance between sensitivity and specificity and yielded predictive performance comparable to that of the best individual classifiers. Nevertheless, only modest improvements were observed over several individual models, and the weighted soft voting ensemble did not consistently outperform the optimized Random Forest classifier.

A comparative evaluation of the ensemble and individual machine learning models is presented in Table 11. Although both ensemble approaches demonstrated satisfactory predictive capability, the optimized Random Forest model remained the best-performing classifier overall, achieving the highest combination of discrimination and classification accuracy. Consequently, the Random Forest model was selected for subsequent model interpretation and diagnostic analyses. The confusion matrix has been shown in Fig 7.

3.6. Feature importance

The five important features that significantly affect the model's predictions are clearly visualized in Fig 6. These features emphasize the qualities most relevant to the prediction of waterborne disease and are essential in determining the model's output. The most significant variable was age, which had a score of 0.27, suggesting that age is a significant factor in determining susceptibility to waterborne disease. This was followed by flood frequency severity (FS) at 0.24 and household size (HS) at 0.14, suggesting that flood severity and household size have a major impact on disease risk. The

Table 11. Comparative performance metrics of the ensemble and individual machine learning models.

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC (95% CI)
Random Forest	0.614	0.719	0.641	0.568	0.678	0.643 (0.519-0.738)
Logistic Regression	0.554	0.694	0.531	0.595	0.602	0.62 (0.511-0.720)
XGBoost	0.604	0.658	0.781	0.297	0.714	0.607 (0.512-0.732)
KNN	0.564	0.628	0.766	0.216	0.69	0.620 (0.525-0.739)
SVM	0.624	0.641	0.922	0.108	0.756	0.596 (0.451-0.682)
Weighted Soft Voting	0.594	0.642	0.812	0.216	0.717	0.636 (0.520-0.728)
Stacking Ensemble	0.594	0.639	0.828	0.189	0.721	0.638 (0.526-0.736)

https://doi.org/10.1371/journal.pwat.0000542.t011

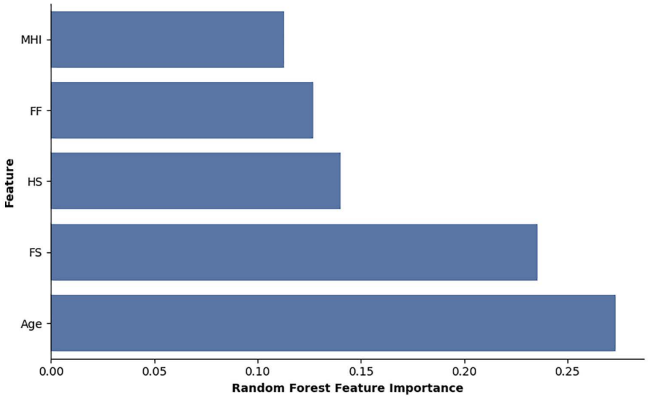


Fig 6. Top five predictors identified by the tuned Random Forest model.

https://doi.org/10.1371/journal.pwat.0000542.g006

significance of environmental factors and socioeconomic status was reflected in the scores of 0.13 and 0.11 for flood frequency (FF) and monthly household income (MHI), respectively. These results highlight how the risk of waterborne disease is influenced by a combination of environmental and demographic factors.

3.7 Model interpretation

To improve the interpretability of the best-performing Random Forest model, SHAP and permutation feature importance analyses were conducted. These analyses identified the predictors that contributed most to the model's classification decisions and illustrated how individual variables influenced the predicted probability of waterborne disease.

3.7.1 SHAP summary plot. The SHAP summary plot was generated to evaluate the global contribution of each predictor to the optimized Random Forest model (Fig 8). The SHAP summary plot provides an interpretation of the contribution of individual features to the model prediction. The features are arranged according to their overall impact on the model output, measured by the mean absolute SHAP value. The x-axis represents the SHAP value, where positive values indicate that a feature increases the predicted outcome, whereas negative values indicate a reduction in the prediction. The colour scale represents the feature magnitude, with red indicating high feature values and blue indicating low feature values.

Based on the SHAP analysis, FS, FF, MHI, HS, and Age demonstrate the largest influence on the model predictions, which is consistent with the feature importance analysis identifying Age, FS, HS, FF, and MHI as the five most influential variables. Although the ranking order differs slightly between the two approaches, both analyses confirm that these variables are the primary contributors to the model decision-making process.

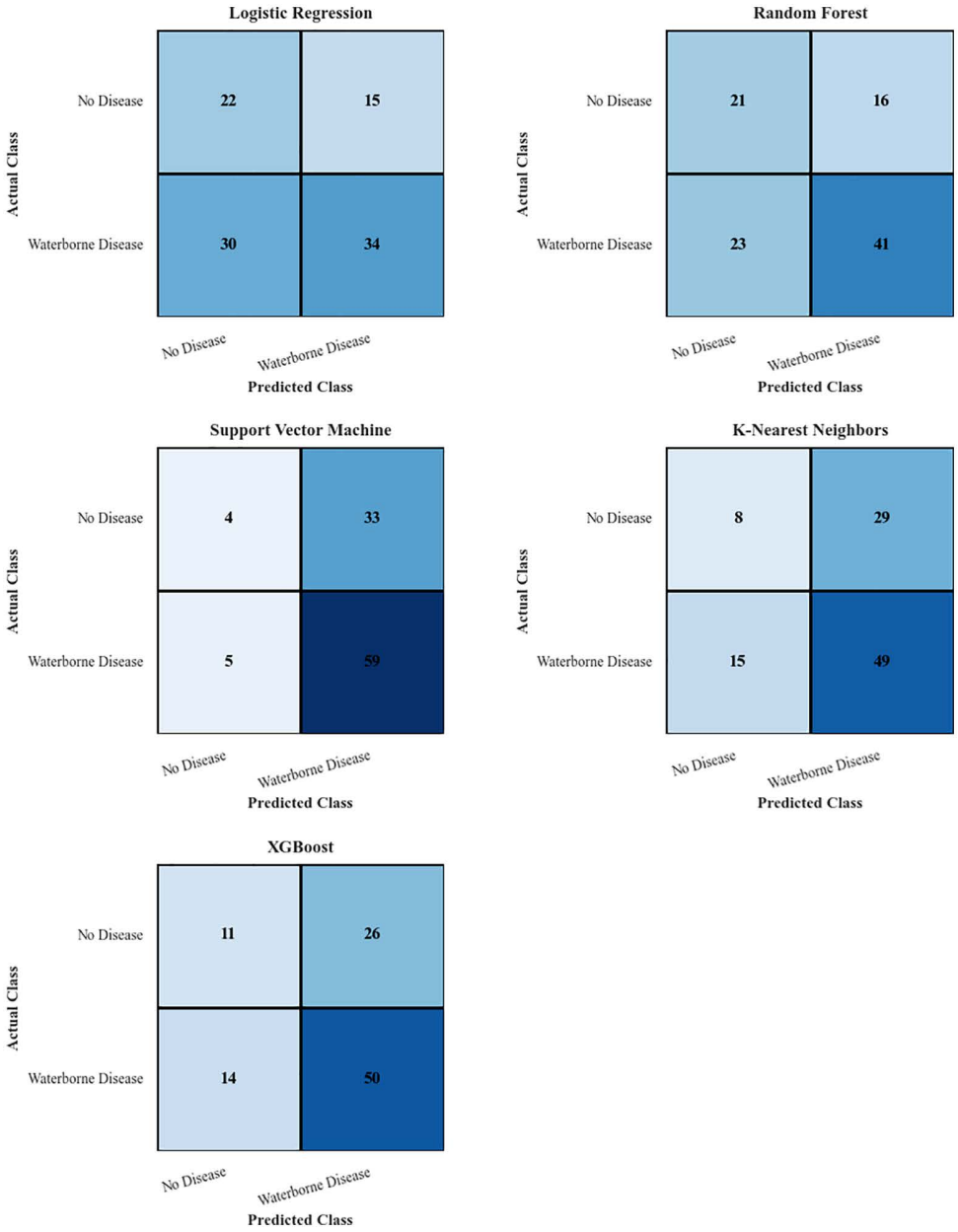


Fig 7. Confusion matrix for the applied classification models.

https://doi.org/10.1371/journal.pwat.0000542.g007

Among these features, FS exhibits the highest SHAP contribution, with values distributed across a wide range from negative to positive impacts. High FS values (red points) are mainly associated with positive SHAP values, indicating that increased FS values tend to increase the model output, while lower FS values contribute negatively. FF also shows a strong influence, where lower feature values are mostly associated with positive SHAP effects, suggesting an inverse relationship between FF and model predictions.

The feature MHI shows a notable contribution, with higher values generally producing positive SHAP effects, whereas lower values tend to decrease the predicted outcome. HS also has a considerable impact, with higher values contributing

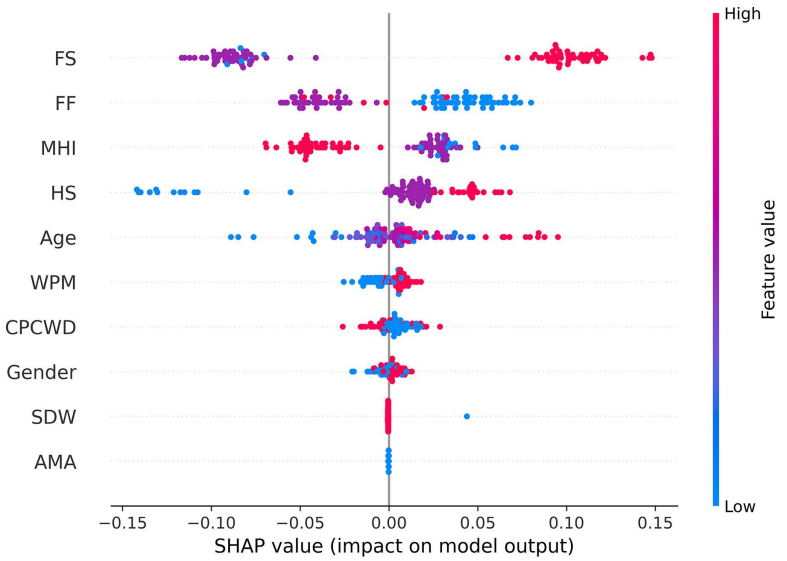


Fig 8. SHAP summary plot showing the impact of features on model prediction.

https://doi.org/10.1371/journal.pwat.0000542.g008

positively and lower values producing negative effects, although some variation is observed. Age, despite having a slightly lower SHAP ranking compared with FS, FF, MHI, and HS in the summary plot, remains one of the most important variables according to the overall feature importance analysis. Its SHAP distribution indicates that age-related variations influence model predictions in both positive and negative directions.

The remaining features, including WPM, CPCWD, Gender, SDW, and AMA, show relatively smaller SHAP distributions concentrated around zero, indicating limited influence on the model output compared with the top five features.

Overall, the SHAP interpretation confirms that Age, FS, HS, FF, and MHI are the key predictive features, highlighting their dominant role in explaining the model behaviour. The consistency between SHAP analysis and feature importance ranking strengthens the reliability and interpretability of the developed model.

This version avoids a contradiction between the feature importance chart and SHAP plot by explaining that ranking differences can occur while the same top predictors remain influential.

3.7.2 SHAP dependance plot analysis. The SHAP dependence plots were used to investigate the relationship between individual input variables and their contributions to the model predictions. The vertical axis represents the SHAP value, where positive values indicate a positive contribution toward the predicted outcome and negative values indicate a reduction in the model output. The color gradient represents the magnitude of an interacting variable, allowing visualization of possible feature interactions.

For FS (Fig 9(a)), the model response shows a clear separation among the categorical FS levels. The lowest FS category is associated with predominantly negative SHAP values, indicating a decreasing influence on the model prediction. In contrast, the highest FS category produces consistently positive SHAP contributions, suggesting that increased FS strongly enhances the predicted response. The narrow distribution within each category indicates relatively stable model behavior, while the color variation (Age) suggests that age may slightly modify the influence of FS.

For FF (Fig 9(b)), the dependence pattern indicates a non-linear relationship. Lower FF values are mainly associated with positive SHAP contributions, whereas intermediate and higher FF categories generally show negative contributions. This suggests that FF has a threshold-dependent effect, where increasing FF beyond a certain level reduces the model prediction. The variation in color (MHI) indicates a possible interaction between household income level and FF, with different MHI values influencing the magnitude of FF effects.

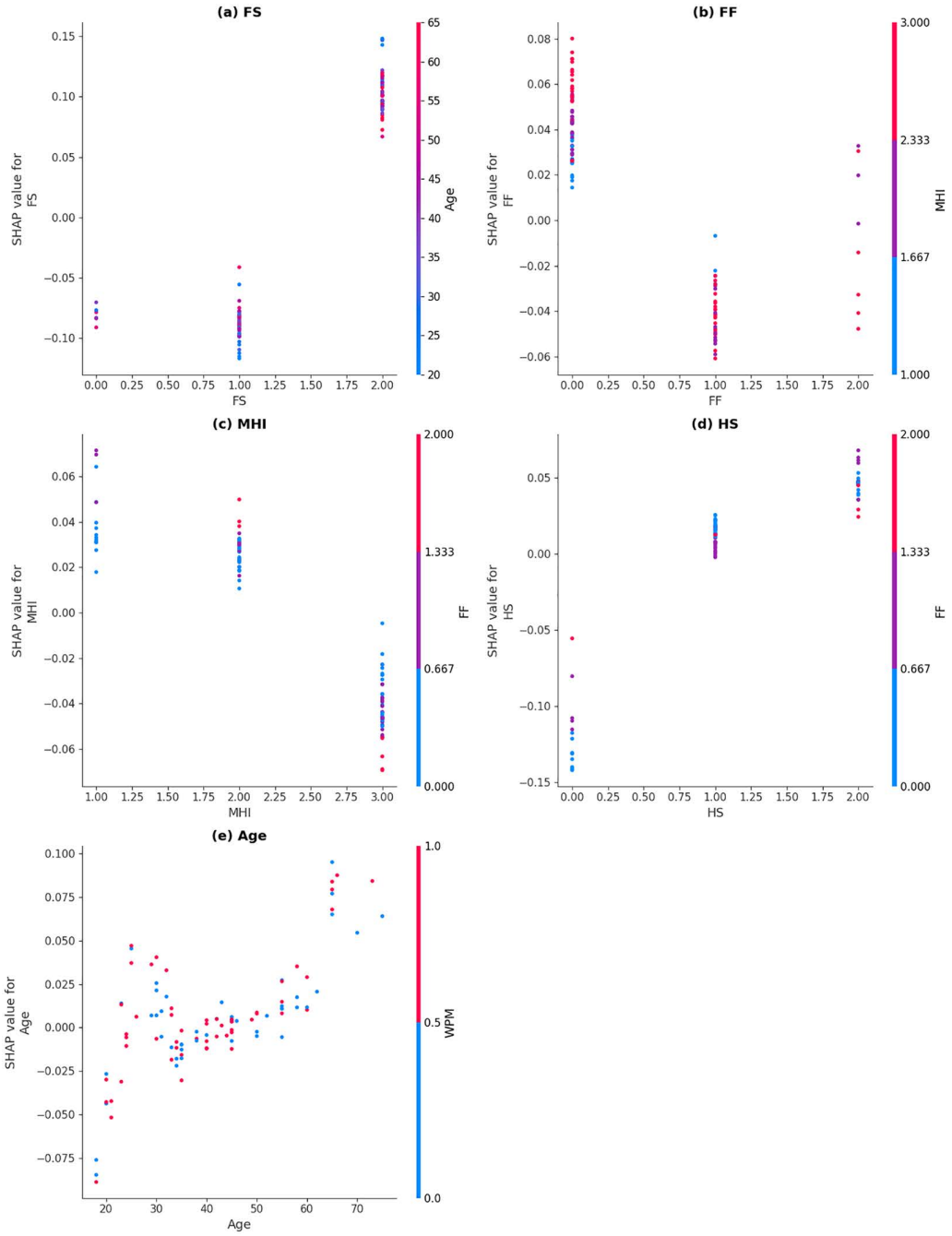


Fig 9. SHAP dependence plots illustrating the influence of key predictors on model output. (a) FS, (b) FF, (c) MHI, (d) HS, and (e) Age.

https://doi.org/10.1371/journal.pwat.0000542.g009

The dependence plot for MHI (Fig 9(c)) demonstrates a strong categorical effect. The lowest MHI category contributes positively to the model output, while higher MHI categories show increasingly negative SHAP values. This indicates that MHI is an important predictor and that changes in socioeconomic conditions significantly affect the model response. The color distribution (FF) suggests that the effect of MHI may be influenced by differences in FF conditions.

For HS (Fig 9(d)), the SHAP distribution shows a progressive shift from negative to positive contributions with increasing HS values. The lowest HS category produces strongly negative SHAP values, while the highest HS category contributes positively. This indicates that HS has a substantial influence on the prediction, with higher HS conditions associated with increased model output. The color gradient (FF) suggests an interaction between HS and FF, although the dominant pattern is controlled by HS itself.

The dependence plot for Age (Fig 9(e)) reveals a continuous and approximately increasing relationship between age and model contribution. Younger ages are associated with mostly negative SHAP values, whereas older individuals show increasingly positive contributions. This indicates that age has a gradual nonlinear effect on the model prediction. The spread of SHAP values at higher ages suggests that other variables, represented by the color scale (WPM), may influence the strength of this relationship.

Overall, the SHAP dependence analysis confirms that FS, FF, MHI, HS, and Age contribute differently to the model predictions, with several variables exhibiting nonlinear effects and potential interactions. Categorical variables (FS, FF, MHI, and HS) demonstrate distinct shifts in model response between categories, whereas Age exhibits a continuous trend. These findings provide additional interpretability by identifying how variations in key predictors influence the direction and magnitude of the model output. The observed SHAP patterns support the importance ranking obtained from the global SHAP analysis and highlight the complex relationships captured by the machine-learning model.

3.7.3 Permutation feature importance. To further evaluate model interpretability and assess the robustness of predictor contributions, permutation feature importance analysis was performed for the Random Forest model. Unlike the intrinsic feature importance derived from tree splitting criteria, permutation importance quantifies the impact of each predictor on model performance by measuring the reduction in ROC-AUC after randomly shuffling the values of an individual feature. Therefore, features with larger decreases in ROC-AUC are considered to have greater predictive relevance.

As presented in Fig 10, the permutation analysis demonstrated that FS showed the largest decrease in ROC-AUC following permutation, indicating that this feature contributed substantially to the predictive performance of the Random Forest model. FF, HS, and MHI also showed relatively higher permutation importance, suggesting that these variables provided additional predictive information. In contrast, variables such as Gender, WPM, and SDW demonstrated smaller effects, while AMA, CPCWD, and Age showed limited or slightly negative changes in ROC-AUC after permutation.

The differences observed between permutation importance and the built-in Random Forest feature importance rankings may be attributed to the distinct principles underlying these approaches. The conventional feature importance measure evaluates the frequency and effectiveness of feature-based splits within individual decision trees, whereas permutation importance reflects the effect of removing a predictor on the overall predictive performance of the trained model. Therefore, a variable such as Age, which showed high importance in the intrinsic feature importance analysis, may contribute through interactions with other predictors or influence tree construction without producing a large independent decrease in ROC-AUC when permuted.

Taken together, both feature importance approaches provide complementary insights into model behavior. The intrinsic feature importance analysis identified Age, FF, and MHI as major contributors to model construction, whereas permutation importance highlighted FS, FF, HS, and MHI as features with the greatest impact on predictive performance. The consistent identification of FF and MHI across both approaches supports their importance as robust predictors, while the differences emphasize the value of applying multiple explainability methods to better understand complex machine-learning models.

3.8 Random forest model diagnostics

3.8.1 Out-of-Bag (OOB) performance. To further assess the internal performance of the optimized Random Forest model, the Out-of-Bag (OOB) score was calculated using the bootstrap samples generated during model training. The

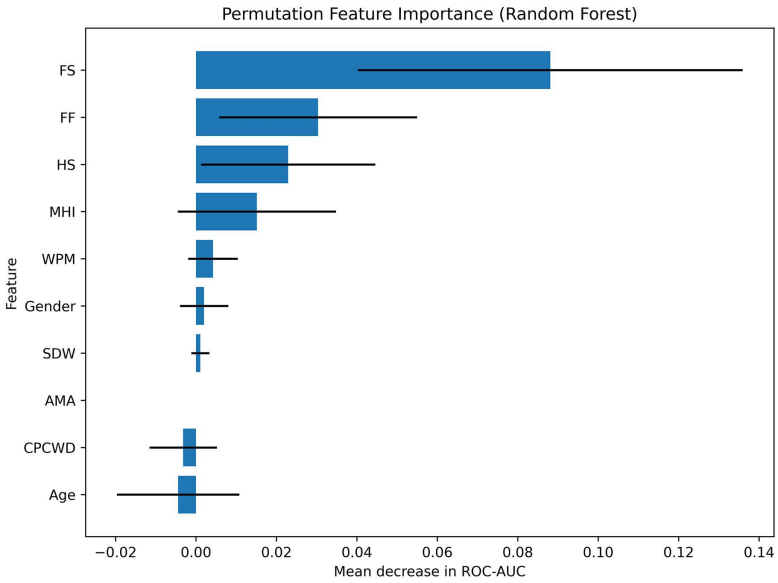


Fig 10. Permutation feature importance analysis of the random forest model.

https://doi.org/10.1371/journal.pwat.0000542.g010

optimized Random Forest model demonstrated an out-of-bag (OOB) accuracy of 0.6125, corresponding to an OOB error rate of 0.3875. The OOB estimate, calculated using bootstrap samples generated during model training, provides an internal assessment of model generalization without requiring an additional validation dataset. Although the OOB accuracy was lower than the accuracy observed on the independent testing dataset, it represents a more conservative estimate of predictive performance because each observation was evaluated only by trees for which it was not included in the bootstrap sample. The OOB results indicate that the Random Forest model maintained reasonable predictive capability during training while exhibiting acceptable generalization performance and no evidence of substantial overfitting. The relatively consistent OOB performance, together with the independent test-set evaluation, supports the robustness of the optimized Random Forest classifier for predicting waterborne disease risk in the study population.

3.8.2 Out-of-bag error curve. The out-of-bag (OOB) error curve for the optimized Random Forest model is presented in Fig 11. The OOB error rate exhibited moderate fluctuations during the early stages of model construction when relatively few trees were included in the ensemble. As the number of trees increased, the error rate gradually stabilized, indicating convergence of the Random Forest model. Beyond approximately 300 trees, only minor variations in the OOB error were observed, suggesting that additional trees provided limited improvement in predictive performance. The final OOB error rate converged to 0.3875, corresponding to an OOB accuracy of 61.25%, demonstrating stable model performance throughout the ensemble-building process. The absence of substantial increases in the OOB error with increasing numbers of trees indicates that the optimized Random Forest model achieved a stable generalization error and did not exhibit evidence of instability or excessive variance. Overall, the convergence of the OOB error curve supports the reliability of the selected Random Forest configuration and confirms that the chosen number of trees was sufficient to achieve stable predictive performance.

3.8.3 Gini feature importance. The Gini feature importance analysis was performed to quantify the contribution of each predictor variable in the Random Forest model based on the mean decrease in impurity (MDI). This metric measures how much each feature contributes to reducing the uncertainty within decision trees, with higher values indicating greater influence on the model's predictive performance. The ranking of feature importance is presented in Fig 12.

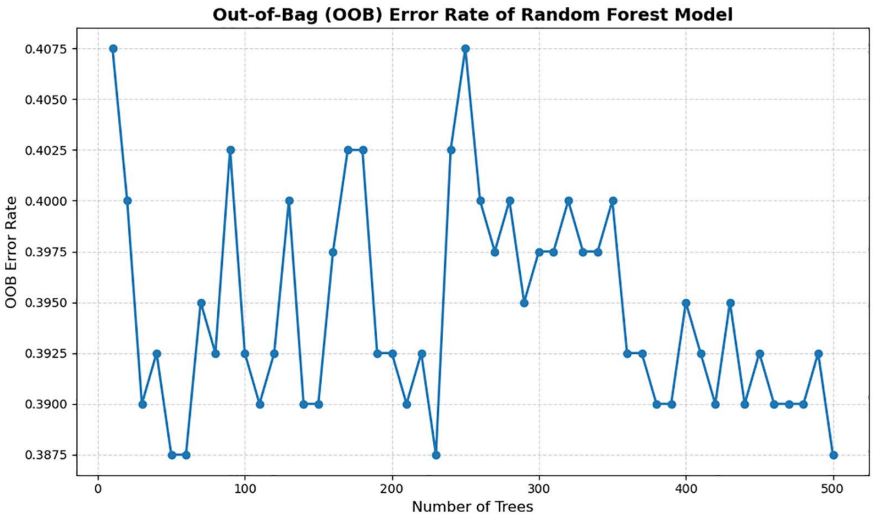


Fig 11. Out-of-Bag (OOB) error curve of the optimized Random Forest model showing the convergence of the OOB error rate as the number of decision trees increased. The stabilization of the error after approximately 300 trees indicates that the ensemble reached a stable level of predictive performance, with a final OOB error rate of 0.3875 (OOB accuracy = 61.25%). Fig 11. Variation of Out-of-Bag Error Rate with Increasing Number of Trees in the Random Forest Model.

https://doi.org/10.1371/journal.pwat.0000542.g011

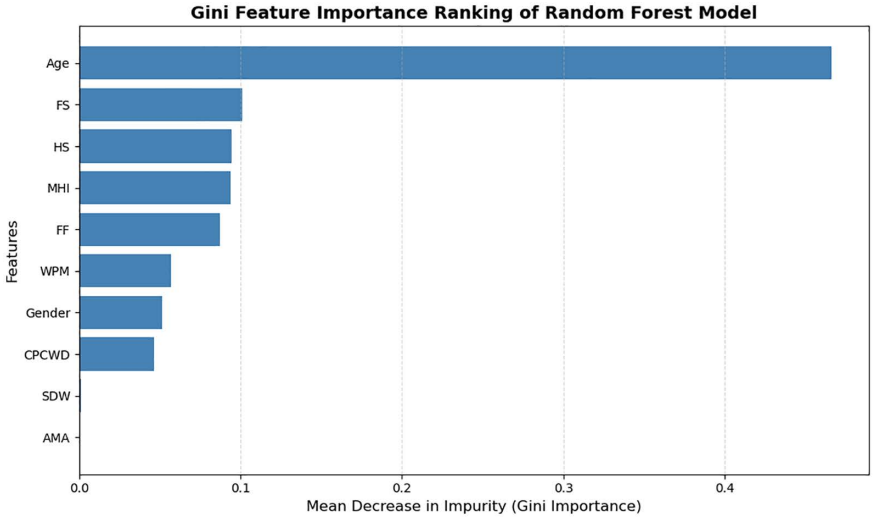


Fig 12. Gini feature importance ranking of the Random Forest model based on mean decrease in impurity.

https://doi.org/10.1371/journal.pwat.0000542.g012

The results demonstrate that Age was the most influential predictor in the Random Forest model, exhibiting a substantially higher Gini importance value (~0.46) compared with all other variables. This indicates that age contributed the greatest reduction in node impurity and played a dominant role in distinguishing the target outcomes. The strong influence of age suggests that variations associated with different age groups are important determinants in the prediction process.

The next most influential variables were FS (~0.10), HS (~0.095), MHI (~0.095), and FF (~0.087), which showed moderate contributions. WPM, Gender, and CPCWD had lower importance values but still contributed to the model's predictive

performance. In contrast, SDW and AMA showed minimal importance, suggesting limited influence on classification outcomes.

Overall, the feature importance ranking indicates that Age was the primary predictor, while FS, HS, MHI, and FF provided additional predictive information in the Random Forest model.

4. Discussion

Waterborne diseases continue to impose a substantial public health burden in low- and middle-income countries, particularly in flood-prone settings where disruption of water, sanitation, and hygiene (WASH) infrastructure increases the risk of pathogen transmission. Bangladesh is especially vulnerable because recurrent flooding, population density, inadequate sanitation, and socioeconomic inequalities collectively facilitate the spread of diarrheal and other waterborne diseases. In this context, accurate identification of individuals at increased risk is essential for implementing timely prevention and response strategies. The present study developed and compared multiple machine learning algorithms to predict waterborne disease among residents of two flood-affected districts in Bangladesh. Among the evaluated classifiers, the optimized Random Forest model achieved the highest predictive performance, demonstrating superior discrimination and calibration compared with Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and XGBoost. Furthermore, model interpretability analyses consistently identified age, flood severity, household size, flood frequency, and monthly household income as the most influential predictors.

The superior performance of the Random Forest algorithm is consistent with its ability to model complex nonlinear relationships and higher-order interactions among demographic, environmental, and socioeconomic variables without requiring strong distributional assumptions. Unlike conventional regression approaches, Random Forest combines multiple decision trees through bootstrap aggregation, thereby reducing variance while maintaining predictive stability. These characteristics are particularly advantageous for epidemiological datasets where multiple correlated risk factors jointly influence disease occurrence. The robustness of the Random Forest model observed in this study is further supported by the SHAP analysis, permutation feature importance, and Out-of-Bag (OOB) diagnostics, all of which demonstrated stable predictor rankings and satisfactory internal model performance.

Our findings are consistent with previous studies reporting superior predictive performance of Random Forest for waterborne and infectious disease prediction. For example, Random Forest achieved higher predictive accuracy for typhoid and malaria than competing algorithms using hospital-based data from Pakistan [27]. Similar findings have been reported for cholera, typhoid, and diarrheal disease prediction, where Random Forest consistently demonstrated high classification accuracy across different study populations [28,29]. Studies conducted in West Africa [30], multi-county cholera surveillance systems [31], Nigeria [32], East Africa [33], and Ethiopia [34,35] likewise identified Random Forest as one of the most accurate machine learning algorithms for predicting waterborne or diarrheal diseases. Although differences in predictive performance across studies are expected because of variations in sample size, disease type, predictor variables, and validation strategies, the consistent superiority of Random Forest across diverse epidemiological settings supports its application as a reliable tool for public health risk prediction.

Beyond predictive performance, the present study provides important insights into the determinants of waterborne disease in flood-affected communities. Feature importance analyses consistently highlighted age, flood severity, household size, flood frequency, and monthly household income as dominant predictors. These findings are biologically and epidemiologically plausible. Age has long been recognized as an important determinant of susceptibility to waterborne infections because younger individuals generally possess less developed immunity and are more vulnerable to dehydration and enteric infections, while older adults may experience increased susceptibility because of underlying chronic conditions and reduced physiological resilience [36,37]. Similarly, greater flood severity and more frequent flooding increase environmental contamination through the overflow of sanitation systems, contamination of drinking water sources, and disruption of hygiene practices, thereby facilitating transmission of waterborne pathogens [38].

Socioeconomic conditions also played a substantial role in disease prediction. Lower monthly household income was associated with increased vulnerability, likely reflecting reduced access to improved drinking water, sanitation facilities, hygiene resources, and healthcare services. Likewise, larger household size may facilitate household-level transmission through overcrowding and increased person-to-person contact. These findings are consistent with previous studies demonstrating that inadequate sanitation, unsafe drinking water, and poor hygiene practices remain major contributors to diarrheal diseases in resource-limited settings [39,40]. Collectively, these results emphasize that waterborne disease risk is influenced not only by environmental exposures but also by socioeconomic vulnerability and household living conditions.

An important strength of this study is the comprehensive evaluation framework used to assess model performance. In addition to conventional discrimination measures, model calibration, bootstrap confidence intervals, DeLong's test, SHAP analysis, permutation feature importance, and OOB diagnostics were incorporated to evaluate predictive performance, robustness, and interpretability. Such comprehensive evaluation remains relatively uncommon in studies applying machine learning to waterborne disease prediction and provides greater confidence in the reliability and practical applicability of the developed model. Furthermore, the use of explainable artificial intelligence techniques enhances transparency by demonstrating how individual predictors contribute to model predictions, thereby improving the potential acceptance of machine learning tools among clinicians, epidemiologists, and public health decision-makers [41].

From a public health perspective, these findings have important practical implications. The optimized Random Forest model could be incorporated into community-based surveillance systems to identify high-risk households before or during flood events, thereby enabling targeted interventions such as emergency water treatment, sanitation support, hygiene promotion, and early healthcare referral. Integrating predictive analytics with existing flood early warning systems could improve resource allocation and strengthen preparedness strategies in disaster-prone regions of Bangladesh. Similar data-driven decision-support systems have been increasingly recommended for climate-sensitive infectious disease surveillance and disaster risk management [42].

Several limitations should be acknowledged. First, the cross-sectional design precludes causal inference and limits assessment of temporal changes in disease risk. Longitudinal cohort studies would provide stronger evidence regarding the dynamic relationships between flooding, environmental conditions, and waterborne disease occurrence. Second, the study population was restricted to two flood-affected districts, which may limit the generalizability of the findings to other geographical regions or countries with different environmental and socioeconomic characteristics. External validation using independent datasets from other flood-prone areas would further strengthen the robustness of the proposed model. Third, although several widely used machine learning algorithms were evaluated, more advanced deep learning and ensemble learning approaches could be explored in future studies to determine whether additional improvements in predictive performance can be achieved. Finally, the inclusion of environmental, meteorological, microbiological, and remotely sensed data may further enhance predictive accuracy and support the development of real-time early warning systems.

In conclusion, the optimized Random Forest model demonstrated excellent predictive performance for identifying individuals at risk of waterborne disease in flood-affected communities of Bangladesh and consistently outperformed the competing machine learning algorithms. The identification of age, flood severity, flood frequency, household size, and monthly household income as key predictors provides valuable evidence for designing targeted prevention and intervention strategies. By combining robust predictive performance with model interpretability, the proposed framework offers a practical and transparent decision-support tool that could assist public health authorities in strengthening waterborne disease surveillance, improving disaster preparedness, and reducing disease burden in flood-prone settings.

5. Conclusion

This study developed and comprehensively evaluated multiple machine learning models to predict waterborne disease among residents of flood-affected communities in Bangladesh. Among the evaluated algorithms, the optimized Random

Forest classifier consistently demonstrated the best overall predictive performance, achieving superior discrimination and calibration compared with Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and XGBoost. The robustness of the model was further supported by bootstrap validation, Out-of-Bag diagnostics, and explainable artificial intelligence techniques, including SHAP and permutation feature importance analyses.

The model interpretation consistently identified age, flood severity, household size, flood frequency, and monthly household income as the most influential predictors of waterborne disease, highlighting the combined importance of demographic, environmental, and socioeconomic factors in disease risk. These findings provide valuable evidence for identifying vulnerable populations and improving risk stratification in flood-prone settings.

The proposed prediction framework has important implications for public health practice. Integrating the optimized Random Forest model into community-based surveillance and disaster preparedness systems could facilitate the early identification of high-risk households, enabling targeted interventions such as safe water provision, sanitation improvement, hygiene promotion, and timely healthcare services. Such data-driven approaches have the potential to improve resource allocation and strengthen preparedness during flood events in Bangladesh and other low-resource settings with similar environmental challenges.

Future research should validate the proposed model using independent external datasets from different geographical regions and incorporate longitudinal, environmental, meteorological, and remotely sensed data to further improve prediction accuracy and generalizability. The evaluation of advanced deep learning and hybrid ensemble approaches may also enhance future prediction systems. Overall, the findings demonstrate that interpretable machine learning, particularly the Random Forest algorithm, provides a reliable and practical decision-support tool for waterborne disease prediction and can contribute to more effective public health surveillance, early warning, and disease prevention strategies in flood-affected communities.

Supporting information

S1 Fig. Pairwise association among predictors by correlation heatmap.
(TIF)

S1 Table. Hyperparameter search space used for RandomizedSearchCV.
(DOCX)

S2 Table. Optimal hyperparameters identified through RandomizedSearchCV using five-fold cross-validation.
(DOCX)

S3 Table. Variance Inflation Factor (VIF) values for candidate predictor variables.
(DOCX)

S4 Table. Pairwise comparison of the discriminatory performance of Random Forest and other classifiers using DeLong's test.
(DOCX)

Acknowledgments

We would like to express our sincere gratitude to the people of Noakhali and Feni for their valuable participation in the survey and interviews. We would also like to thank the students of the Department of Statistics, Noakhali Science and Technology University, who served as interviewers and contributed their time and effort to the field-level data collection. We further acknowledge the Research Cell, Noakhali Science and Technology University, Noakhali, Bangladesh, for providing financial support for this research.

Author contributions

Conceptualization: Md. Mamun Miah.

Data curation: Md. Mamun Miah, Kabir Hossain, H.M. Shahadat Ali.

Formal analysis: Md. Mamun Miah, Kabir Hossain, H.M. Shahadat Ali.

Investigation: Md. Mamun Miah, Hafiz T.A. Khan.

Methodology: Md. Mamun Miah.

Project administration: Md. Mamun Miah.

Resources: Md. Mamun Miah.

Software: Md. Mamun Miah.

Supervision: Md. Mamun Miah, Hafiz T.A. Khan.

Validation: Md. Mamun Miah.

Visualization: Kabir Hossain, Hafiz T.A. Khan, H.M. Shahadat Ali.

Writing – original draft: Md. Mamun Miah, Kabir Hossain.

Writing – review & editing: Md. Mamun Miah.

References

1. World Health Organization WHO. Bangladesh. Accessed 2025 May 15. https://cdn.who.int/media/docs/default-source/searo/wsh-och-searo/bangladesh-wsp.pdf
2. Chanda Shimi A, Ara Parvin G, Biswas C, Shaw R. Impact and adaptation to flood. Disaster Prevent Manag: An Int J. 2010;19(3):298–313. https://doi.org/10.1108/09653561011052484
3. Waterborne disease. Accessed 2025 May 15. https://www.undrr.org/understanding-disaster-risk/terminology/hips/bi0018
4. World Health Organization WHO. Drinking water. Accessed 2025 May 15. https://www.who.int/news-room/fact-sheets/detail/drinking-water
5. World Vision. Waterborne disease facts and how to help. Accessed 2025 May 15. https://www.worldvision.ca/stories/clean-water/cholera-waterborne-disease-facts
6. Institute for Health Metrics and Evaluation (IHME). GBD compare. 2019. Accessed 2025 May 15.
7. Chowdhury AH, Rahman MS. Machine learning and spatio-temporal analysis of meteorological factors on waterborne diseases in Bangladesh. PLoS Negl Trop Dis. 2025;19(1):e0012800. https://doi.org/10.1371/journal.pntd.0012800 PMID: 39820842
8. WaterAid B. Bangladesh - facts and statistics. Accessed 2025 August 17.
9. Akter M, Kabir H. Health inequalities in rural and urban Bangladesh: the implications of digital health. Mayo Clin Proc Digit Health. 2023;1(2):201–2. https://doi.org/10.1016/j.mcpdig.2023.04.003 PMID: 40206725
10. Dubal ZB, Kolhe R, Das S, Rawat S. Waterborne diseases. Management of fish diseases. Singapore: Springer Nature Singapore; 2025. 119–34. https://doi.org/10.1007/978-981-96-0270-4_4
11. Sedhain P. Water, sanitation, socioeconomic status and prevalence of waterborne diseases: a cross-sectional study at Makwanpur district, Nepal. 2014. https://munin.uit.no/handle/10037/6503
12. Akanbi O. Poverty as a precursor for spatial pattern of waterborne diseases prevalence in Kwara state, Nigeria: a geographer's view. Ethiop J Env Stud Manag. 2016;9(1):45. https://doi.org/10.4314/ejesm.v9i1.5
13. Walker J. The influence of climate change on waterborne disease and Legionella: a review. Perspect Public Health. 2018;138(5):282–6. https://doi.org/10.1177/1757913918791198
14. Rahaman MA, Kalam A, Al-Mamun M. Unplanned urbanization and health risks of Dhaka City in Bangladesh: uncovering the associations between urban environment and public health. Front Public Health. 2023;11. https://doi.org/10.3389/fpubh.2023.1269362
15. Grampurohit S, Sagarnal C. Disease prediction using machine learning algorithms. In: 2020 International Conference for Emerging Technology (INCET), 2020. 1–7. https://doi.org/10.1109/incet49848.2020.9154130
16. Lyimo NN, Fue KG, Materu SF, Kilatu NC, Telemala JP. Assessing the potential for application of machine learning in predicting weather-sensitive waterborne diseases in selected districts of Tanzania. Front Artificial Intelligence. 2025;8. https://doi.org/10.3389/frai.2025.1597727
17. Hussain M, et al. Machine learning based efficient prediction of positive cases of waterborne diseases. BMC Med Inform Decis Mak. 2023;23(1):11. https://doi.org/10.1186/s12911-022-02092-1

18. Campbell AM, Racault M-F, Goult S, Laurenson A. Cholera risk: a machine learning approach applied to essential climate variables. Int J Environ Res Public Health. 2020;17(24):9378. https://doi.org/10.3390/ijerph17249378
19. Chowdhury FS, Bin Zaman S, Mahmood SAI. Access to water and awareness about the unsafe water in rural Bangladesh. J Med Res Innov. 2017;2(1):e000088. https://doi.org/10.15419/jmri.88
20. Niu L. A review of the application of logistic regression in educational research: common issues, implications, and suggestions. Educ Rev. 2020;72(1):41–67. https://doi.org/10.1080/00131911.2018.1483892
21. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sensing. 2005;26(1):217–22. https://doi.org/10.1080/01431160412331269698
22. Almatarneh S, Gamallo P, Pena FJR, Alexeev A. Supervised classifiers to identify hate speech on english and spanish tweets. 2019. 23–30. https://doi.org/10.1007/978-3-030-34058-2_3
23. Wang L, Khan L, Thuraisingham B. An effective evidence theory based k-nearest neighbor (KNN) classification. In: 2008 IEEE/WIC/ACM International conference on web intelligence and intelligent agent technology. 2008. 797–801. https://doi.org/10.1109/WIIAT.2008.411
24. Zhang P, Jia Y, Shang Y. Research and application of XGBoost in imbalanced data. Int J Distrib Sens Netw. 2022;18(6):155013292211069. https://doi.org/10.1177/15501329221106935
25. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024;14(1):6086. https://doi.org/10.1038/s41598-024-56706-x PMID: 38480847
26. Faraggi D, Reiser B. Estimation of the area under the ROC curve. Stat Med. 2002;21(20):3093–106. https://doi.org/10.1002/sim.1228 PMID: 12369084
27. Hussain M, et al. Machine learning based efficient prediction of positive cases of waterborne diseases. BMC Med Inform Decis Mak. 2023;23(1):11. https://doi.org/10.1186/s12911-022-02092-1
28. Taral SV. A Comparative review of predictive models for waterborne disease outbreaks. Int J Adv Comp Eng Commun Tech. 2025;14(1):278–86. https://doi.org/10.65521/ijacect.v14i1.455
29. Taral SV, Vishnoi Y. Machine learning-based classification of waterborne diseases. Int J Adv Comp Eng Commun Tech. 2025;14(1):184–93.
30. Onyijen OH, Olaitan EO, Olayinka TC, Oyelola S. Data-driven machine learning techniques for the prediction of cholera outbreak in West Africa. Western Euro J Modern Exp Scientific Methods. 2023;1(1):33–51.
31. Dattangire R, Biradar D, Vaidya R, Joon A. A comprehensive analysis of cholera disease prediction using machine learning. 2025. 555–68. https://doi.org/10.1007/978-981-96-2697-7_39
32. B. OC, Omankwu EE. Leveraging machine learning for early detection and prediction of cholera outbreaks in Nigeria: a data-driven approach. The Transac Nigerian Assoc Mathematical Physics. 2024;20:73–82.
33. Yehuala TZ, Derseh NM, Tewelgne MF, Wubante SM. Exploring machine learning algorithms to predict diarrhea disease and identify its determinants among under-five years children in East Africa. J Epidemiol Glob Health. 2024;14(3):1089–99. https://doi.org/10.1007/s44197-024-00259-9
34. Kassaw A, Alebachew Muluneh A, Assefa EM, Yimer A. Predictive modeling and socioeconomic determinants of diarrhea in children under five in the Amhara Region, Ethiopia. Front Public Health. 2024;12. https://doi.org/10.3389/fpubh.2024.1366496
35. Zemariam AB, Abey W, Kassaw AK, Yimer A. Comparative analysis of machine learning algorithms for predicting diarrhea among under-five children in Ethiopia: evidence from 2016 EDHS. Health Informatics J. 2024;30(3). https://doi.org/10.1177/14604582241285769
36. Hasan MN, Siddiqui MN, Akter MstF, Mitu S, Chowdhury MAB, Uddin MJ. Assessing contributory factors of diarrhea among under-five children in Bangladesh from 2006 to 2019 and recent increases: a cross-sectional study. Health Sci Rep. 2025;8(2). https://doi.org/10.1002/hsr2.70457
37. Sarker AR, Sultana M, Mahumud RA, Sheikh N, Van Der Meer R, Morton A. Prevalence and health care–seeking behavior for childhood diarrheal disease in Bangladesh. Glob Pediatr Health. 2016;3. https://doi.org/10.1177/2333794X16680901
38. Wang P, Asare EO, Pitzer VE, Dubrow R, Chen K. Floods and diarrhea risk in young children in low- and middle-income countries. JAMA Pediatr. 2023;177(11):1206. https://doi.org/10.1001/jamapediatrics.2023.3964
39. Birhan TA, Bitew BD, Dagne H, Amare DE, Azanaw J, Genet M, et al. Prevalence of diarrheal disease and associated factors among under-five children in flood-prone settlements of Northwest Ethiopia: a cross-sectional community-based study. Front Pediatr. 2023;11:1056129. https://doi.org/10.3389/fped.2023.1056129 PMID: 36756237
40. Yirdaw G, Mekonen H, Assaye BT, Amare GA, Yenew C. Prevalence of acute diarrhea and its risk factors among under five children in flood affected Dasenech District, Southern Ethiopia: a cross-sectional study. Sci Rep. 2025;15(1):16980. https://doi.org/10.1038/s41598-025-02120-w PMID: 40374893
41. Sadeghi Z, Alizadehsani R, CIFCI MA, Kausar S, Rehman R, Mahanta P, et al. A review of Explainable Artificial Intelligence in healthcare. Computers and Electrical Engineering. 2024;118:109370. https://doi.org/10.1016/j.compeleceng.2024.109370
42. Alhumaidi NH, Dermawan D, Kamaruzaman HF, Alotaiq N. The use of machine learning for analyzing real-world data in disease prediction and management: systematic review. JMIR Medical Inform. 2025;13(1): e68898.