



An optimized data analytics pipeline for improving healthcare diagnosis using ensemble learning

Lomat Haider Chowdhury^a, Shaira Tabassum^b, Swakkhar Shatabda^c, Ashir Ahmed^d

^a Ahsanullah University of Science and Technology, Dhaka, Bangladesh

^b Norwegian University of Science and Technology, Trondheim, Norway

^c BRAC University, Dhaka, Bangladesh

^d Kyushu University, Fukuoka, Japan

ARTICLE INFO

Keywords:

Healthcare diagnosis
Data analytics
Noise handling
Portable Health Clinic
COVID-19

ABSTRACT

Healthcare diagnosis is a process physicians follow before prescribing the patients. The medical doctors may make an early prediction by observing the physical signs and symptoms. Imposing a treatment without proper diagnosis cannot guarantee a cure and sometimes may lead the patient to a more detrimental scenario. However, the cost of healthcare diagnosis makes people indifferent to going through the process. Big data and machine learning are already in use to contribute to the healthcare diagnosis sector with the available data which is enormously growing through the digitalization of the system. Yet the difficulty remains since the raw data contains noise including missing values, outliers, and an imbalanced number of samples. These properties in a dataset make it challenging to implement any diagnosis model. A complete patient profile cannot be generated due to missing values, which may affect the final prediction. Outliers in a medical dataset represent extreme cases and rare conditions, or they may even be generated due to data entry errors. An excessive number of outliers may lead to a skewed and incorrect prediction. An imbalanced dataset makes it challenging to identify the minority classes appropriately and mostly generates a biased model for majority class instances. A combination of advanced preprocessing techniques and reliable model selection are required to address these challenges effectively. This paper proposes a data analytics pipeline on a Portable Health Clinic (PHC) dataset. The paper systematically evaluates different preprocessing methods for missing value imputation, outliers detection, and data balancing and offers a comprehensive preprocessing framework. Later, five state-of-the-art ensemble models for healthcare diagnosis were implemented along with a proposed ensemble machine learning model, KNN-XGBoost-SVM-Random Forest (KNN-X-SVM-R). The proposed model achieved an accuracy of 97.03% which supersedes all the other state-of-the-art models. To reaffirm the rectification of our model, we experimented with it on another COVID-19 routine blood test dataset. In both cases, our proposed model acquired better results regarding different performance measures. Validating the approach on a secondary dataset strengthens the robustness of the proposed methodology. The recommended preprocessing and modeling approach can be adopted to enhance diagnostic systems and improve patient outcomes.

1. Introduction

With advancements in the application of machine learning models, numerous fields have achieved considerable developments in the last few years. Machine learning has benefited branches like economics, medicine, robotics, business analytics, computing, and many more. It is being used in forecasting weather conditions [1], autonomous navigation and artificial intelligence in Robotics [2], market trend analysis from stock market data [3], mental health analysis from social interactions [4], and so on.

Healthcare is one of the industries that have used this advancement for its growth and is a very demanding field for research. At the same time, the field is also quite sensitive and challenging since the data source must be reliable and quality must be ensured, the privacy and security of the data should be maintained. Moreover, the healthcare and medical data are heterogeneous in nature. Machine learning research and models for healthcare and medicine can be used to predict disease and possible treatments, analyze medical imaging and diagnosis, determine the effectiveness of new vaccines and medicines,

* Corresponding author.

E-mail address: lomathaider.cse@aust.edu (L.H. Chowdhury).

<https://doi.org/10.1016/j.imu.2025.101623>

Received 28 August 2024; Received in revised form 23 January 2025; Accepted 28 January 2025

Available online 5 February 2025

2352-9148/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

group similar disease symptoms and their impacts, and organize medical records. The outbreak of COVID-19 has given researchers in medical industries many more opportunities to contribute and determine the possible actions [5], diagnosis, treatments, vaccines [6,7], and post-vaccination outcomes of coronavirus [8,9]. Significant development has been achieved in diverse medical sectors such as partial annotations for large medical image structures using deep learning to optimize the analysis cost [10], classification of lung diseases from CT scan images using Support Vector Machine (SVM) approach [11] and feature extraction for Alzheimer disease using convolutional neural network (CNN) [12].

According to the report of Statista Research Department, the United States has the highest per capita health expenditure of 12,318 U.S. dollars which is more than 17% of the total gross domestic product (GDP) [13]. The report also claims that the percentage may increase to 20% by 2028 [14]. A large number of healthcare costs depend on the diagnosis of health problems. In the era of machine learning and deep learning methods, we plan to improve healthcare diagnosis to reduce the overall expenses for each patient. These models may not replace the activities of medical doctors, however, these may help both patients and physicians. With the enormous health monitoring devices, researchers now have a large amount of data available to apply machine learning models for health diagnosis. However, this raw data contains ample noise to provide the prospective outcomes. In this study, we proposed a method to reduce the noise in the data for a better diagnosis result.

While applying a machine learning model, one must follow some steps for reliable outcomes. The basic steps are Data Collection and Integration, Data Preprocessing, Model Selection, and Implementation. In this work, different data analytics pipelines were implemented. The best-performing method for missing value handling, outliers detection, and sampling of the data has been determined. For imputing missing values, a total of eight imputation methods have been implemented that include Listwise deletion, Fast KNN, KNN imputer, Data wig, Multiple imputations by chained equation (MICE), using Median, Most frequent, and Mean. Similarly, to detect and remove outliers, Z-score, IQR score, iForest, Local outlier factor (LOR), Elliptic, One class, Stochastic gradient descent (SGD), and DBscan have been used. To balance the dataset, Oversampling, Undersampling, and Synthetic minority oversampling techniques (SMOTE) were implemented. The data analytics methods are depicted in Fig. 1. An optimal preprocessing method for healthcare datasets has been suggested through experiments.

Sometimes, applying only one existing model may not provide the required result or outcome, even can be overfitted. In that case, an ensemble approach can be applied where a collection of models helps to get a better prediction. For instance, in article [15], the researchers proposed a Rider Optimization Algorithm-Neural Network (RideNN) and Deep Neuro Fuzzy Neural Network (DNFN) for diagnosis prediction of Coronary Heart Disease. Research [16] proposed a hybrid model combining Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) to predict cardiac abnormality from ECG signals. For this work, we propose an ensemble learning approach using a Portable Health Clinic (PHC) dataset. Then, the performance of the proposed model KNN-XGBoost-Support Vector Machine-Random Forest (KNN-X-SVM-R) has been compared to an existing ensemble model Extra Trees-Random Forest-Logistic Regression-XGBoost (ERLX) that was used for the COVID-19 routine blood test dataset discussed in this article [17]. Later, the result of our proposed model is validated on the COVID-19 dataset. Validation on a secondary dataset improves the credibility of the research to use in diverse healthcare scenarios. In both cases, our proposed model performed better according to accuracy, F1-score, Specificity, and other evaluation measures.

2. Related work

The research field of machine learning in healthcare is widely popular. Several kinds of research have been done in the last few decades with some alluring outcomes.

Studies from 2017 to 2019 highlighted the uses of different supervised and unsupervised machine learning models in clinical contexts. Callahan et al. in [18] present the advantages of using machine learning over other conventional analysis techniques. The article focuses on the applications of machine learning models in the healthcare sector to achieve maximum potential. It also represents the challenges and limitations of using machine learning models. Research [19] discussed different supervised, semi-supervised, unsupervised, and reinforcement learning models that can be used to analyze these diverse healthcare data.

Research has been done following Fast Healthcare Interoperability Resources (FHIR) standards to align the healthcare data with Health Level 7 (HL7). Articles [20,21] worked on interoperability of healthcare data. Specification similarity, syntactic similarity, and semantic similarity have been measured for data transformation to provide a generic data acquisition API for several health-related factor identifiers such as BP monitor, Activity tracker, Pulse Oximeter, Body weight, and Glucometer. Depending on the urgency and computational needs of medical scenarios, Internet of Medical Things (IoMT) device data and network requirements are segmented into separate 5G network slices where each slice transforms the data to resolve heterogeneity of data. Another study [22] integrates the Personal Health Record (PHR) system with Electronic Health Records (EHR) to maximize the outcome of patient and doctor communication. They serve a bi-directional and almost real-time exchange of information between PHR and EHR.

With time, the research focus shifted to deep learning models from linear machine learning techniques with the availability of data and enough computation resources. In [23], Ahmad et al. discusses the various applications for interpretable machine learning and deep learning models in healthcare. The interpretability of model predictions is noted by clinical providers and other healthcare decision-makers as a priority for implementation and utilization. According to the authors, the important factors that affect the explanation are input data, model parameters, the algorithms used, user's competency, domain knowledge (novice vs. expert), level of explanation (explanation granularity), etc. The study asserts that a deep learning model utilizing straightforward intuitive features may be more interpretable than a simple linear model using highly engineered and sophisticated features.

Miotto et al. annotated the challenges in applying feature engineering to complex, heterogeneous, and unstructured biomedical data to identify more robust features and apply learning models [33]. Later, the authors discussed the highly used deep learning methods to overcome this problem and suggested comprehensive and meaningful architectures to link human interpretability and deep learning models for next-generation healthcare. The authors mainly focused on the differences between traditional machine learning models [34] and traditional Artificial Neural Networks (ANNs) with the deep learning approach [35]. Finally, it represented how different deep learning models perform better on diverse biomedical data than the other traditional models. Article [36] discusses different deep-learning techniques in computer vision (e.g. medical imaging), natural language processing (e.g. electronic health record data), and reinforcement learning (e.g. robotic-assisted surgery).

In [24], authors used four healthcare datasets of kidney disease, heart disease, diabetes, and breast cancer. After preprocessing the data, the authors suggested a feature selection model using Support Vector Machine (SVM) and Recursive Feature Elimination (RFE). For hyperparameter optimization, the Bayesian optimization method has been suggested. Finally, XG-Boost was used to optimize the model. The model accuracy varies from 81% to 100% for different datasets.

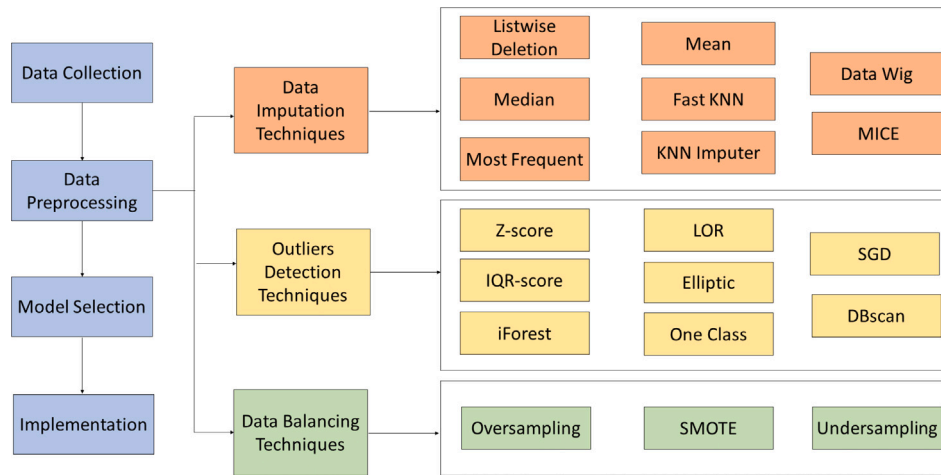


Fig. 1. Different Data Preprocessing Techniques.

Table 1

A summary of related works from the healthcare domain.

Article	Data	Methods
Shailaja et al. [19]	Diabetes, Breast Cancer, Thyroid	J48, SVM, Naive Bayes, Decision Tree
Ahmed et al. [23]	Pneumonia	Interpretable Machine Learning, Explainable AI
Al-Jamimi et al. [24]	Kidney Disease, Heart Disease, Diabetes, Breast Cancer	Feature Selection: Recursive Feature Elimination (RFE); Hyperparameter: Bayesian Optimization; Model: XG-Boost
Raza et al. [25]	Heart Disease	Feature Selection: Particle Swarm Optimization (Majority Voting); Ensemble Model: Bagged Trees+Random Forest+AdaBoost
Kiourtis et al. [26], Kouremenou et al. [27]	Healthcare data	Transformation of data into HL7 FHIR structure through syntactic and semantic similarities
De mollo et al. [28]	Virtuoso, Neo4J, HBase, ARM, Jena TBD, BaseX, Oracle, Think!EHR, PostgreSQL	Systematic review on solving interoperability. Emphasizes semantic interoperability in EHR to improve healthcare system integration
Verma et al. [29]	Breast Cancer, Dermatology, Hepatitis, Liver-disorder, Splice	Bagging: AdaBoost; Boosting: Random Forest; Stacking: J48, Naive Bayes, Random Forest
Misra et al. [30]	PIMA Indian Diabetes Dataset, LUGB2	Missing value imputation: KNN; Data scaling: Min-max scaler; Feature selection: Random Forest; Model: Logistic Regression, Random Forest, ANN, SVM
Alim et al. [31]	Heart Disease	Correlation analysis: Pearson; Normalization: Min-Max; Model: Random Forest with K-fold cross-validation
Al Jame et al. [17]	COVID-19	Missing value imputation: KNN Imputer; Outliers removal: iForest; Data balancing: SMOTE; Ensemble Model: Extra Trees+Random Forests+Logistic Regression
Saihood et al. [32]	Pima Indians Diabetes	Missing value imputation: Mean Imputation; Outliers detection: IQR; Data balancing: SMOTE; Normalization: Min-Max; Feature Selection: RFE; Bagging: Random Forest; Boosting: XG-Boost; Stacking: Random Forest+SVM

Ensemble learning is another widely used approach that implements a set of machine learning models for better analysis. Azad et al. suggests a multi-layered ensemble learning model for binary classification [37]. In article [25], authors used ensemble learning with a majority voting rule to predict heart disease from previously stored medical data (UCI ML Heart Disease Dataset). In feature selection, Particle Swarm Optimization (PSO) was used to remove attributes with the lowest ranking score. The authors combined Bagged Trees, Random Forest, and AdaBoost for the ensemble learning model. This ensemble model gained 88.88% accuracy which is better than some other previous models

applied to the same dataset. Research [29] offered another ensemble learning approach which is Bagging, Boosting, and Stacking (BBS) for classification problems. The authors used AdaBoost for Bagging, Random Tree for Boosting method, and J48, Naive Bayes, and Random Forest for Stacking. Their model was implemented on five different bioinformatics datasets and achieved better accuracy and lower mean squared error for each.

Along with model implementation, data processing also plays an important role in better intuition of the data and result analysis. Several research has been done to offer useful data processing methods. Alim

et al. proposed a prediction model for heart disease using the UCI vascular heart disease dataset in [31]. The article mainly focused on identifying the highly correlated feature that can maximize the prediction result. The researchers followed the steps to prepare the model, which included missing value imputation, outlier detection, correlation analysis, normalization, model selection, and implementation. The Min-Max normalization technique was used for normalization. For model implementation, they used the Random Forest model with K-fold cross-validation. However, the methods followed in earlier steps are not mentioned. The research article that we mostly focus on in our study is an Ensemble Learning approach to the COVID-19 dataset [17]. Here, the authors combined different machine learning models and proposed an ERLX model that performed very well for the dataset. To start, they used the KNN Imputer for missing value imputation, the Isolation Forest (iForest) method to remove the outliers, and Synthetic Minority Oversampling Technique (SMOTE) for data balancing. In the second stage, the model combines Extra Trees, Random Forests, and Logistic Regression for ensemble learning. Later Meta algorithm XGBoost was used.

Saihood et al. proposes a three-step model experimenting on Pima Indians Diabetes Database [32]. The very first stage was data preprocessing. Mean imputation, IQR, SMOTE, Min-Max Normalization, and Recursive Feature Elimination were used for missing value imputation, outliers detection, data balancing, normalization, and feature selection respectively. However, the article does not address how they chose the abovementioned methods. Hyperparameter was selected using the Grid-search technique. Random Forests and XG-Boost were used for Bagging and Boosting as ensemble models. For stacking, Random Forests and Support Vector Machines were stacked.

Healthcare data are often complex, error-prone, heterogeneous, inconsistent, and challenging. The challenges in this domain are discussed in [38]. The authors suggested validation, cleaning, verification, and logging for data cleaning mechanisms to improve the reliability and quality of the data. The mechanism comprises three sub-components for ingesting, storing, and cleaning data through four key actions: Validation, Cleaning, Verification, and Logging. Using established techniques, the mechanism was evaluated on three healthcare datasets with varying errors, demonstrating its effectiveness and efficiency in producing high-quality, reliable data.

The research that has been done on the healthcare sector is quite vast. A summary of relevant research is presented in Table 1. Recent developments in machine learning and deep learning have enlarged the research scope to a great extent. However, there is a lack of a complete data analytics pipeline on healthcare diagnosis. In some cases, the dataset is not robust enough to assert the reliability of the result. In this work, we propose an ensemble learning model with a data analytics pipeline used on reliable healthcare data and achieved an outstanding result.

3. Healthcare datasets

This section describes the datasets used for the proposed model. Data is vital when applying an algorithm or a model and analyzing the problem. In this work, two datasets have been used, one is the Diagnosis of COVID-19 dataset and another one is a Portable Health Clinic (PHC) dataset.

3.1. COVID-19 dataset

This dataset is collected from the Albert Einstein Israelita Hospital of Saulo Paulo, Brazil which is publicly available in Kaggle [39]. The data was collected from 28 March 2020 to 3 April 2020, during the period when the COVID-19 affection rate was at its peak. The dataset is used in research where an ensemble learning model is proposed for COVID-19 diagnosis [17]. There are 5,644 samples and the dataset contains 111 attributes having more than 100 diagnosis tests including

urine tests, mineral element tests, blood tests, Adenovirus tests, rt-PCR tests, SARS-CoV-2 tests, and so on. Among the features, the SARS-CoV-2 tests are majorly responsible for COVID-19 diagnosis. Out of the 111 attributes, research [17] used 18 features that are responsible for COVID by applying the feature selection method. The features are Hemoglobin, Platelets, Leukocytes, Lymphocytes, Basophils, Eosinophils, Neutrophils, Monocytes, Patient Age, Urea, C-reactive Protein, Creatinine, Potassium, Sodium, Alanine transaminase, Aspartate transaminase, International normalized ratio (INR), Albumin. Table 2 represents the statistical observations about the features, available data, missing values, and the rate of missing data. The table depicts that out of the total 5,644 samples only the feature 'Patient age quantile' is complete. The other features contain a notable amount of missing values. The minimum rate of missing values is 4.70% for the features of Hemoglobin, Platelets, Leukocytes, Lymphocytes, Basophils, Eosinophils, and Monocytes. The maximum rate of missing values is 5.25% for the feature Albumin. There are 1,07,236 data values out of which only 18,680 data are available and the rest 88,556 data points are missing. Thus, the missing data rate becomes 82.58% which is huge. Though the researchers suggest an ensemble model in [17] for the COVID diagnostic test, this large amount of missing data will hamper any proposed model and provide a biased result. The predicted result from such a dataset cannot be implemented in real life even if we apply the best missing value imputation techniques since the medical sector is very sensitive. This limitation inspires us to choose a better dataset with fewer missing values and more stability.

3.2. PHC dataset

The Portable Health Clinic or PHC dataset is collected from the rural areas of Bangladesh containing the health information for 36,462 entries with 18 different health-related attributes. Based on the features the target is to identify the health status of a particular person whether it is suspicious, affected, or emergent. The overview of the PHC dataset is depicted in Table 3 which represents that apart from Gender, Age, and Health Status, the Blood glucose feature contains the least missing values which are 0.40%. The maximum percentage of missing values is for the features Urinary glucose and protein, 1.92%. There are 6,56,316 data points where 5,26,251 values are present and only 1,30,065 values are missing. Therefore, the rate of missing values in the PHC dataset is 19.82% which is very commendable compared to the COVID-19 dataset that had an 82.58% missing value rate. Different research has already been done with the PHC dataset such as [40–42]. While preparing the dataset, patients agreed to use their data for analysis purposes only. Thus, the PHC dataset is not uploaded to any public repository. The dataset has been reviewed and approved by the Kyushu University Institutional Review Board (Approval No. 20182007).

Considering the rate of missing values, the PHC dataset is preferred over the COVID-19 dataset since the models implemented for this one will be more stable and the analysis will be more accurate to implement in real-life scenarios. However, for the implementation, both datasets have been used for our model and a comparative analysis is discussed.

4. Analysis methodology

This section discusses the comparative analysis between the model of research [17] and our proposed ensemble models. The COVID-19 dataset described in Section 3 was used to implement the prior model. The article used different methods for data preprocessing. For missing value imputation, the authors suggested the KNN Imputer technique. For outliers detection and handling, iForest was suggested. For sampling, SMOTE was used. These are the three steps for data preprocessing. The article used an ensemble approach at two levels to implement the model. At the first level, Extra Trees, Random Forest, and Logistic Regression were used parallelly. In the second layer, the XGBoost model was implemented to boost the performance of the

Table 2
Overview of missing values in the COVID dataset.

SL.	Feature	Available data	Missing data	Missing rate in column (%)	Missing rate in dataset (%)
1	SARSCov	5,644	0	0	0
2	Hemoglobin	603	5,041	89.33	4.70
3	Platelets	602	5,042	89.33	4.70
4	Leukocytes	602	5,042	89.33	4.70
5	Lymphocytes	602	5,042	89.33	4.70
6	Basophils	602	5,042	89.33	4.70
7	Eosinophils	602	5,042	89.33	4.70
8	Neutrophils	513	5,131	90.91	4.78
9	Monocytes	601	5,043	89.35	4.70
10	Patient age quantile	5,644	0	0	0
11	Urea	397	5,247	92.97	4.89
12	Proteina C reativa	506	5,138	91.03	4.79
13	Creatinine	424	5,220	92.49	4.87
14	Potassium	371	5,273	93.43	4.92
15	Sodium	370	5,274	93.44	4.92
16	Alanine transaminase	225	5,419	96.01	5.05
17	Aspartate transaminase	226	5,418	96.00	5.05
18	Normalized ratio	133	5,511	97.64	5.14
19	Albumin	13	5,631	99.77	5.25
Total		18,680	88,556	–	82.58

Table 3
Overview of missing values in the PHC dataset.

SL.	Feature	Available data	Missing data	Missing rate in column (%)	Missing rate in dataset (%)
1	Gender	36,462	0	0	0
2	Age	36,455	7	0.02	0
3	Height	26,323	10,139	27.81	1.54
4	Weight	27,100	9,362	25.68	1.43
5	BMI	27,435	9,027	24.76	1.38
6	Waist	24,588	11,874	32.57	1.81
7	Hip	24,588	11,874	32.57	1.81
8	Waist hip ratio	25,127	11,335	31.09	1.73
9	Temperature	25,276	11,186	30.68	1.70
10	Blood oxygen	25,107	11,355	31.14	1.73
11	Blood pressure (sys)	33,318	3,144	8.62	0.48
12	Blood pressure (dia)	33,309	3,153	8.65	0.48
13	Blood glucose	33,822	2,640	7.24	0.40
14	Urinary glucose	23,888	12,574	34.49	1.92
15	Urinary protein	23,868	12,594	34.54	1.92
16	Pulse rate	31,756	4,706	12.91	0.72
17	Arrhythmia	31,367	5,095	13.97	0.78
18	Health status	36,462	0	0	0
Total		526,251	130,065	–	19.82

model and determine the final prediction. Fig. 2(a) represents the block diagram of the existing model [17].

In our proposed methodology, we used a different dataset. The PHC dataset was used with a different implementation technique. After that, the model's performance has been compared with the existing one. At the data preprocessing phase, this article uses MICE over KNN Imputer for missing value imputation; the Z-score model in place of iForest for outliers detection; and SMOTE for data sampling which is similar. For model implementation, this work also follows an ensemble learning approach with different model combinations. At the first layer, KNN, XGBoost, and SVM are proposed. At the second layer, the Random Forest Classifier is used and the final output is predicted. For the implementation, different combinations of models have been tried, and came up with the proposed one performed best for both the COVID-19 dataset and the PHC dataset. In Fig. 2(b), the block diagram of our proposed model has been depicted. The method used for sampling was similar in both cases. However, the rest of the preprocessing techniques for missing value imputation and outliers detection including the ensemble model architecture are different. More detailed discussion regarding the architecture has been done in Section 5 along with the implementation and performance comparison between the models.

5. Ensemble learning model generation

The implementation has been done in three phases. In the first phase, the preprocessing methods for missing value imputation, outlier detection, and sampling were implemented. Six ensemble model combinations have been executed to finalize the model. For this phase, the PHC dataset was used. To check the robustness, existing and proposed models have been implemented on both the PHC and the COVID-19 datasets in the second phase. Later, a comparative analysis has been done. For the last phase, the PHC dataset was used on five state-of-the-art ensemble models previously suggested for healthcare diagnosis along with the proposed method, and a comparative study was done. Fig. 3 depicts the outline of these phases.

Before proposing the model, different methodologies for each phase were tried and figured out the best-performing one. Fig. 4 shows the step-by-step approach from data preprocessing to model implementation procedure for our model. At first, different missing value imputation methods were used. After that, several outliers detection techniques were tried and finally, different data balancing methods were implemented. In every step, the best one was selected and implemented for our proposed model. A combination of ensemble models has

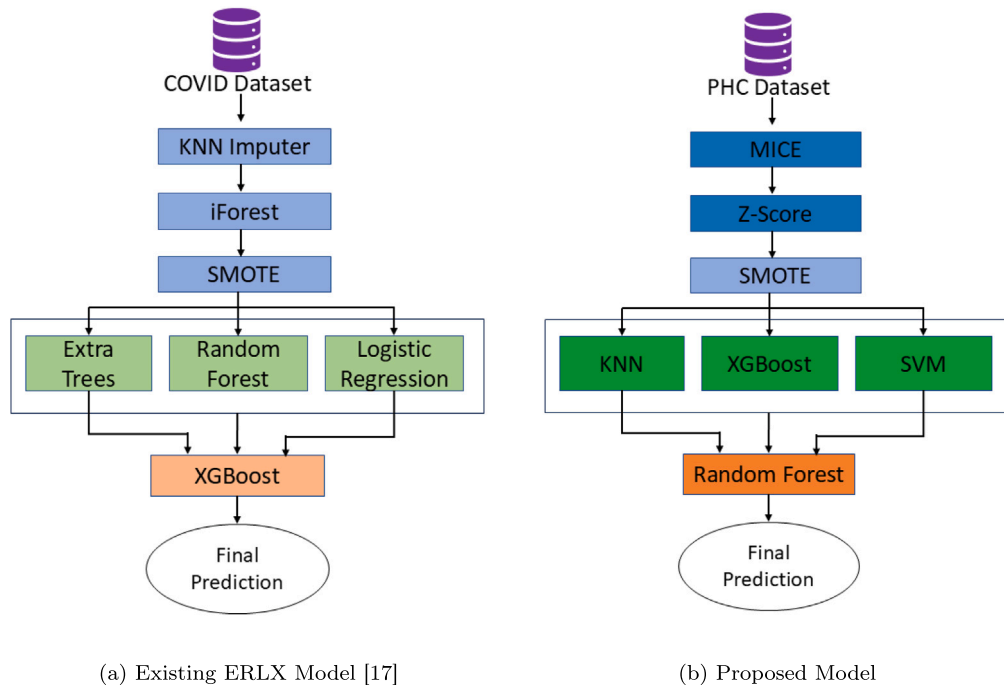


Fig. 2. Block diagram of the existing ERLX model and the proposed model.

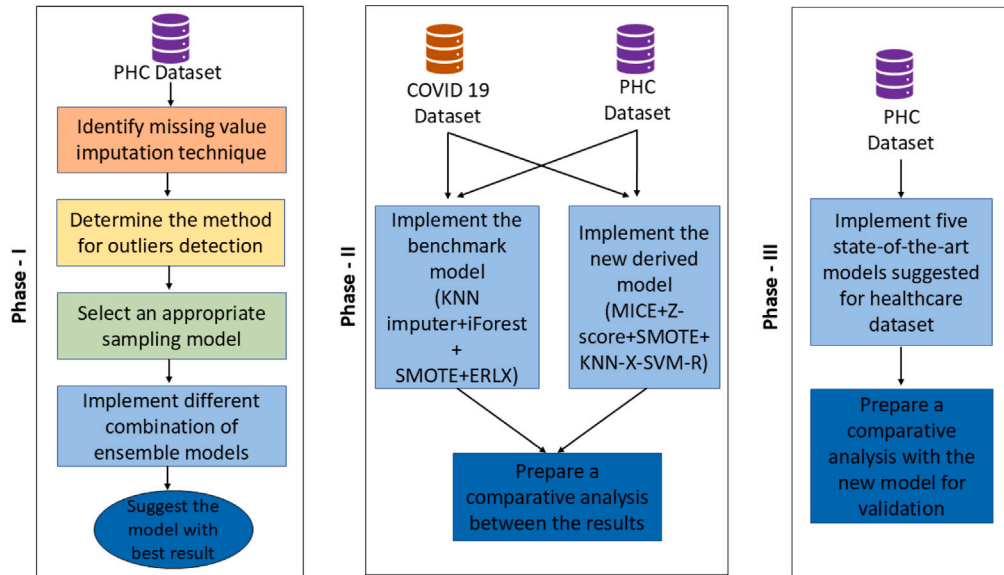


Fig. 3. An outline of the methodology followed throughout the implementation.

also been implemented and described and their performance measures have been observed.

5.1. Missing value imputation methods

The collected raw data may contain some entries for which one or more attribute values are not given. These are called missing values. If the number of missing values is large, the performance of the model may degrade. The easiest way to handle the missing entries is simply dropping the samples. However, this is not an efficient approach since it may reduce the size of the dataset and many important observations may get deleted. To avoid such issues, researchers suggest filling in the missing values with relevant data. There are multiple techniques to fill in the missing values. The most popular missing value imputation

techniques are Listwise Deletion, Median filling, Most Frequent, Mean, Fast KNN, MICE, Data wig, and KNN Imputer. This article denotes them as I1, I2, I3, I4, I5, I6, I7, and I8 respectively.

The existing model ERLX [17] proposed KNN Imputer as their preferred missing value imputation method. From Section 3, we already know that the COVID-19 dataset contains an ample amount of missing values. Therefore, the PHC dataset has been used and the ERLX method was implemented with eight different missing value imputation methodologies (I1-I8). From the observation noted in Table 4, we notice that MICE performed the best in handling missing values compared to the KNN Imputer method for the PHC dataset in terms of Accuracy, AUC, Precision, Recall, F1 score, Sensitivity, and Specificity. The accuracy shows that Listwise deletion has the lowest accuracy score with 92.48%, KNN Imputer achieves 94.07%, whereas MICE has 95.61%

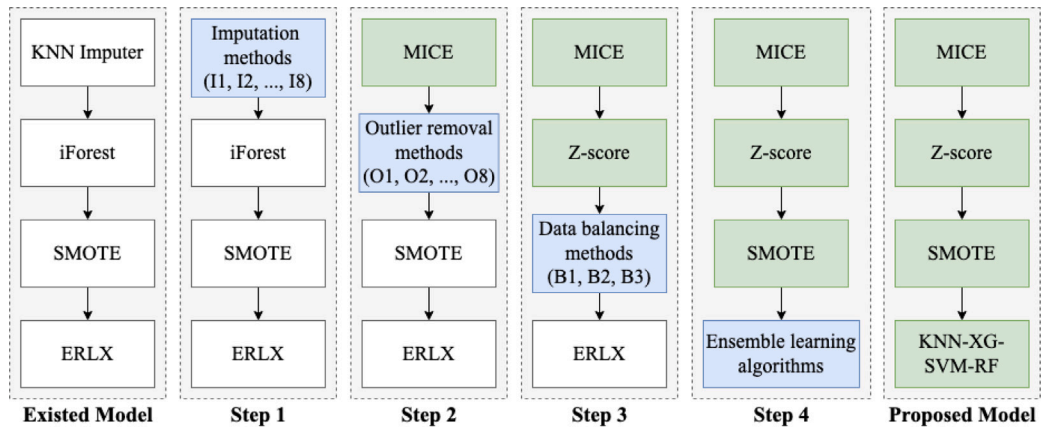


Fig. 4. Step-by-step approach to preprocessing and generating an ensemble learning model.

Table 4

Missing value imputation techniques on the PHC dataset.

Evaluation metrics	I1	I2	I3	I4	I5	I6	I7	I8
Accuracy	92.48	95.43	94.07	93.57	94.51	95.61	95.06	94.07
AUC	90.29	94.43	93.18	92.94	94.11	94.97	94.83	93.67
Precision	92.85	96.00	95.90	95.68	96.65	97.39	96.54	95.88
Recall	80.90	88.90	86.86	85.63	87.69	89.93	89.75	87.42
F1 score	86.46	93.32	91.15	90.38	91.95	93.51	93.02	91.46
Sensitivity	80.90	88.90	86.86	85.63	87.69	89.93	89.75	87.42
Specificity	97.37	98.35	97.98	97.89	98.31	98.69	98.14	97.86

Table 5

Outlier removal techniques with MICE (I6) on the PHC dataset.

Evaluation metrics	O1	O2	O3	O4	O5	O6	O7	O8
Accuracy	94.38	91.24	92.04	92.59	92.87	93.21	93.32	93.38
AUC	95.06	88.80	90.57	91.66	92.27	92.80	93.01	93.06
Precision	94.71	78.26	83.41	86.32	87.96	89.21	90.08	90.52
Recall	90.41	77.75	81.09	83.38	84.60	85.71	86.16	86.25
F1 score	92.51	77.96	82.19	84.79	86.21	87.40	88.05	88.31
Sensitivity	90.41	77.75	81.09	83.38	84.60	85.71	86.16	86.25
Specificity	96.85	94.64	95.37	95.74	95.95	96.17	96.32	96.42

accuracy which is the highest. Thus, the observation declares MICE as the missing value imputation technique for our proposed model.

5.2. Outlier removal methods

The data may contain some noise that needs to be processed for a better prediction model. These noises are called outliers. The outliers can bias the prediction of a model which results in an under-fitting problem. One of the most common ways to identify the outliers in a dataset is by preparing a box plot. There are several methods for outlier detection and removal such as Z-score, IQR-score, iForest, LOR, Elliptic, One class, SGD, and DBscan. This article denotes them as O1, O2, O3, O4, O5, O5, O6, O7, and O8 respectively. Here, the best outliers detection and removal method will be selected using a similar approach used earlier. At this step, the ERLX model on the PHC dataset was implemented for 8 different above-mentioned outliers detection techniques changing the missing value imputation method with MICE, and the performance is represented in Table 5. According to the result, the Z-score method performed best for Accuracy, AUC, Precision, Recall, F1 score, Sensitivity, and Specificity whereas the ERLX model used iForest. The accuracy of the Z-score is 94.38% which is higher than that of iForest which is 92.04%. From this intuition, using a Z-score is proposed for outlier removal for our model.

Table 6

Data balancing techniques with MICE (I6) and Z-score (O1) on the PHC dataset.

Evaluation metrics	Oversampling	SMOTE	Undersampling
Accuracy	93.61	93.68	93.39
AUC	96.48	94.33	95.83
Precision	93.37	94.41	92.96
Recall	90.06	89.32	89.89
F1 score	91.68	91.8	91.39
Sensitivity	90.06	89.32	89.89
Specificity	95.90	96.54	95.65

5.3. Data balancing

Data balancing is a technique of sampling all the input samples with equal distribution labels. This preprocessing technique ensures that the implemented model is not biased toward one label in particular. For example, if a medical dataset contains information mostly from healthy samples, the implemented model will be more biased toward healthy persons. To overcome this problem, properly balancing the data is important. Some data balancing techniques are Oversampling, SMOTE, and Undersampling.

We implemented these three techniques on the PHC dataset with the previous best preprocessing techniques MICE for data imputation from 5.1 and Z-score for outliers removing from 5.2 and then used the ERLX method. Table 6 represents the performance using different performance measures for all the data balancing techniques. An interesting behavior is noticed in the observations. For some measures, Oversampling performed best, for others, SMOTE performed best. For example, Oversampling achieved the best score regarding AUC, Recall, and Sensitivity. Whereas, SMOTE performed better for performance measures like Accuracy, Precision, F1-score, and Specificity. Undersampling performed the least compared to the rest. From this intuition, SMOTE has been considered the best-performing data balancing method as suggested for the COVID-19 dataset in [17].

5.4. Ensemble algorithms combination

At this point, we have the three best-performing data preprocessing techniques: MICE, Z-score, and SMOTE for missing value imputation, outliers removal, and data balancing respectively. In this section, six different ensemble models have been implemented and determined the performance using the preprocessing techniques suggested for the COVID-19 dataset in [17].

Hyperparameters for the models were selected using a manual search method. Table 7 represents the combination of hyperparameters selected for the implemented models. At first, the existing ERLX (Extra Trees, Random Forest, Logistic Regression, XGBoost) model was

Table 7
Hyperparameter settings for different models.

Model	Number of trees	Solver	Max depth	Class weight	Iteration	Multiclass	Learning rate	Neighbors	Kernel	Gamma
Random Forest	300	–	17	balanced	–	–	–	–	–	–
Extra Trees	300	–	17	balanced	–	–	–	–	–	–
Logistic Regression	–	saga	–	balanced	500	One-vs-rest (ovr)	–	–	–	–
XGBoost	300	–	17	–	–	–	1.0	–	–	–
KNN	–	–	–	–	–	–	–	7	–	–
SVM	–	–	–	–	–	–	–	–	RBF	auto

implemented on the PHC dataset. For the Extra Trees and Random Forest classifiers, the number of trees (n_estimators) was 300, and the maximum depth for each tree (max_depth) was 17. For Logistic Regression, ‘saga’ was used as a solver for the optimization problem since the dataset size is large. The maximum number of iterations is kept at 500 for the solver to converge. For the binary class label ‘ovr’ and to adjust the weights for the class labels a ‘balanced’ mode was used that aligns the weight as reversely proportional to class frequencies. Then, the XG-Boost Classifier was implemented with a learning rate of 1.0 for the same number of trees and depth size. The hyperparameters remain the same for any combination of implemented models. This model shows an accuracy of 94.07% with KNN Imputer, iForest, and SMOTE as preprocessing methods whereas 95.61% with our proposed preprocessing methods MICE, Z-score, and SMOTE. This is the first observation where it has been noticed that our proposed preprocessing combination performed better.

For the second ensemble model, we followed two levels of model implementation. In the first layer, Gaussian Naive Bayes (GNB) model is used. In the second layer, a combination of Random Forest, Logistic Regression, and XG Boost are implemented. Thus, the model is named as GNB-RLX model. The hyperparameters for Random Forest, Logistic Regression, and XG-Boost remained the same. Gaussian Naive Bayes (GNB) was implemented with no hyperparameter. The accuracy for this model is 95.28% with our proposed preprocessing technique. On the other hand, the existing data preprocessing combination acquired a lower accuracy of 94.69%.

Extra Trees, Gaussian Naive Bayes (GNB), Random Forest, Logistic Regression, and XG Boost are implemented sequentially for the next model. The model’s name is E-GNB-RLX. The hyperparameters are similar to before. The accuracy with KNN Imputer, iForest, and SMOTE is 94.43% and with MICE, Z-score, and SMOTE is 94.86%.

For the fourth model, a new algorithm Support Vector Machine is introduced along with Gaussian Naive Bayes (GNB), Random Forest, and XG-Boost classifier. A pipeline by standardizing the features with StandardScaler is constructed and the kernel coefficient for the Support Vector Classifier ‘auto’ was chosen, thus, it simply uses the number of instances to determine the coefficient. The model accuracy is 93.94% and 94.86% for existing preprocessing and our proposed preprocessing methods accordingly. The model is named GNB-R-SVM-X. The following implemented model is similar, with a different combination sequence. First, GNB is used, and then XG-Boost, SVM, and Random Forest (GNB-X-SVM-R) are implemented. The sequence of Random Forest and XG-Boost have been swapped and a different accuracy measure is noticed. This time, the accuracy is 95.01% and 96.13% respectively for existing preprocessing and proposed preprocessing methods.

For the last model, another classification algorithm is used which is K-Nearest Neighbors (KNN) instead of GNB. The ensemble model is KNN-X-SVM-R. The model accuracy is 95.32% for KNN Imputer, iForest, and SMOTE, and 97.03% for MICE, Z-Score, and SMOTE as preprocessing techniques.

From the observations in Table 8, it can be inferred that the lowest accuracy for the existing preprocessing methods which is KNN Imputer for missing values, iForest for Outlier removal, and SMOTE for data balancing is 93.94% for the model GNB-R-SVM-X. Whereas, the minimum accuracy with our proposed preprocessing technique which is

Table 8

Ensemble learning model selection on the PHC dataset comparing existing method and proposed method.

Machine learning algorithms	KNN Imputer + iForest + SMOTE	MICE + Z-score + SMOTE
ERLX	94.07	95.61
GNB-RLX	94.69	95.28
E-GNB-RLX	94.43	94.86
GNB-R-SVM-X	93.94	94.86
GNB-X-SVM-R	95.01	96.13
KNN-X-SVM-R	95.32	97.03

Table 9

Comparative analysis between existing method and the proposed method on Covid and PHC datasets.

Evaluation metrics	KNN Imputer + iForest + SMOTE + ERLX		MICE + Z-score + SMOTE + KNN-X-SVM-R	
	Covid dataset	PHC dataset	Covid dataset	PHC dataset
Accuracy	99.88	94.07	99.88	97.03
AUC	99.38	93.67	100	97.90
Precision	100	95.88	100	96.58
Recall	98.72	87.42	98.80	97.44
F1 score	99.36	91.46	99.39	97.00
Sensitivity	98.72	87.42	98.80	97.44
Specificity	99.99	97.86	100	96.64

MICE for missing values, iForest for Outlier removal, and SMOTE for data balancing is 94.86% for ensemble models E-GNB-RLX and GNB-R-SVM-X. Similarly, for any implemented ensemble models, our proposed preprocessing combination performed best. The model with the highest accuracy is K-Nearest Neighbor, XG-Boost, Support Vector Machine, and Random Forest, thus, KNN-X-SVM-R with an accuracy of 97.03%. From the observation, this can be considered the best-implemented ensemble model for medical diagnosis data. The further experiments of this article have been continued following this model.

6. Results analysis

From the previous implementation and discussion, the best-performing ensemble approach is K-Nearest Neighbor, XG-Boost, Support Vector Machine, and Random Forest (KNN-X-SVM-R). Some other experiments have been done using the method to ensure the robustness of the model.

In the earlier discussion, it has been mentioned that paper [17] proposed KNN Imputer, iForest, and SMOTE for preprocessing and ERLX model to diagnose COVID from routine blood tests for the COVID-19 dataset. To analyze the result, the same model is implemented for both the COVID-19 dataset and the PHC dataset. Then, to prepare a comparative analysis, our proposed preprocessing methods MICE, Z-score, and SMOTE are implemented along with the suggested KNN-X-SVM-R model on COVID-19 and the PHC dataset. Table 9 shows the detailed results for both techniques in terms of different performance measures like Accuracy, AUC, Precision, Recall, F1 score, Sensitivity, and Specificity. The result represents that the COVID-19 dataset

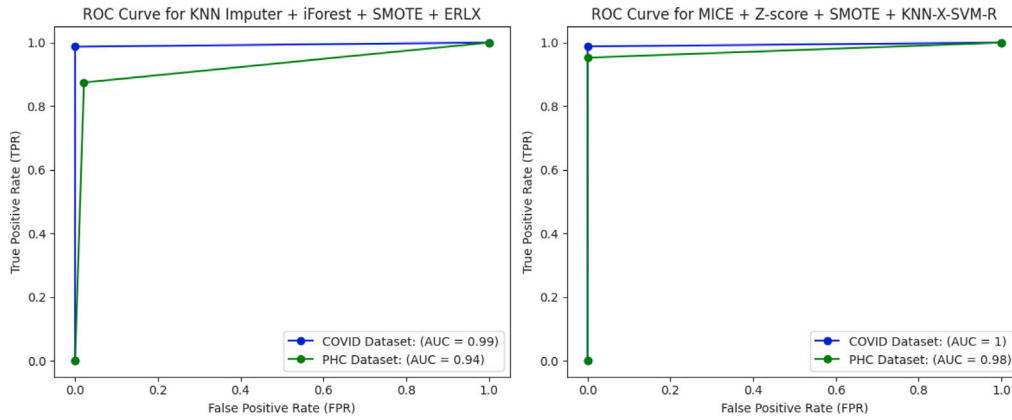


Fig. 5. The ROC curve for COVID-19 and PHC dataset using different models.

achieved a better result for any combination of implementation than the PHC dataset. The “Precision” measure for this dataset is 100% in both cases. Moreover, the dataset achieved a 100% AUC and Specificity for our proposed model. The reason behind the more accurate result could be a large number of missing values (rate: 82.58%) presented in the records that have been described in Section 3. Because of this property, any implemented model shows an overfitted result on the COVID-19 dataset. Whereas, overfitting is not a problem for the PHC data record, where the rate of missing values is 19.82% only. The ROC curve for both the COVID-19 dataset and the PHC dataset has been depicted for the ERLX model and the proposed model in Fig. 5. The graph represents that the area under the curve (AUC) for both datasets is comparatively larger for our proposed model than the previous one. Moreover, our main focus was on the PHC dataset and the AUC has notably improved for this after using the model we offer.

A notable difference has been identified in the results of Recall and Sensitivity between the two models for the PHC dataset. The recall is related to the positive instances of the data thus, the model can perfectly identify healthy instances. Meanwhile, specificity is a performance measure that is related to negative instances of data. Thus, if the model identifies a person as unhealthy is actually unhealthy. Our proposed model achieved a Recall and Sensitivity of 97.44% which is 10.02% more than the existing model ERLX. From this intuition, it can be declared that the proposed KNN-X-SVM-R model performed better along with the preprocessing techniques MICE, Z-score, and SMOTE for both the COVID-19 and the PHC datasets.

To further assert the robustness of the new model, we compared the performance with another four state-of-the-art ensemble models offered for medical diagnosis apart from the ERLX model. In total, five research articles and their models are considered for this comparative analysis. The first one is [29] where a novel ensemble learning approach has been proposed as Bagging-Boosting-Stacking (BBS model). The second one, the ERLX model has been discussed throughout the paper. The next model is proposed in [31] where the Random Forest classifier model has been used with a stratified k-fold cross-validation method. Before applying the model, different data preprocessing methodologies have been followed. In [25], Logistic Regression, Naive Bayes, and Artificial Neural Networks (ANN) have been ensembled with a majority voting technique. The last article is [43] used a feature subset selection method Particle Swarm Optimization (PSO) with an ensemble model for Bagged Trees, Random Forest, and AdaBoost. These five state-of-the-art models have been implemented for the PHC dataset with the proposed K-Nearest Neighbor, XG-Boost, Support Vector Machine, and Random Forest (KNN-X-SVM-R) model. The performance measures are enlisted in Table 10.

The results represent that the proposed KNN-X-SVM-R performed best for any performance measures. Model [31] achieved the lowest accuracy, precision, recall, and F1-score among all the implemented

Table 10

Performance evaluation between the proposed method and existing state-of-the-art ensemble models on the PHC dataset.

Models	Accuracy	Precision	Recall	F1-score
Verma et al. [29]	93.47	93.47	93.29	93.37
Al Jame et al. [17]	94.07	93.11	95.02	94.06
Alim et al. [31]	80.12	78.22	82.77	80.43
Raza et al. [25]	90.13	89.18	91.05	90.11
Yekkala et al. [43]	93.24	91.49	95.15	93.28
Proposed method	97.03	96.58	97.44	97.00

models where the stratified k-fold cross-validation method was proposed with the Random Forest classifier. Other than the KNN-X-SVM-R method, the ERLX model [17] has given a better performance of 94.07%. However, the best accuracy is 97.03%, precision 96.58%, recall 97.44%, and F1-score 97.00% have been achieved using the proposed model K-Nearest Neighbor, XG-Boost, Support Vector Machine, and Random Forest (KNN-X-SVM-R).

Fig. 6 represents the confusion matrices for the above-mentioned models implemented on the PHC dataset. From the values of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), it can be inferred that our proposed model could identify both healthy and unhealthy instances better compared to other state-of-the-art models.

7. Conclusion

With a large amount of available data, healthcare is rapidly approaching machine-driven screening and diagnosis. Data is playing the most significant role here. Along with the enormous opportunities, it also brings challenges to efficiently processing this data to maximize the outcome. Dataset selection is another challenging part of medical diagnosis. Incomplete and noisy data with a large amount of missing values can nullify the robustness of any experiment. Concerning the issue, two datasets have been used in this study. The first one, the COVID-19 dataset has higher missing values in number. The result represents that higher missing values can tend the model to overfit with a 100% precision. Thus, the PHC dataset has been chosen with fewer missing values to minimize the overfitting problem.

In this research, different preprocessing methods for data imputation, outlier detection, and sampling have been practiced. Several combinations of these preprocessing methods have been applied and the best one has been chosen. A similar approach is followed for model implementation. A group of ensemble models has been implemented to determine the model that performs best for the PHC dataset where K-Nearest Neighbor, XG-Boost, Support Vector Machine, and Random Forest (KNN-X-SVM-R) models have been ensembled. Then, the model's performance is cross-validated with the COVID-19 dataset. In both

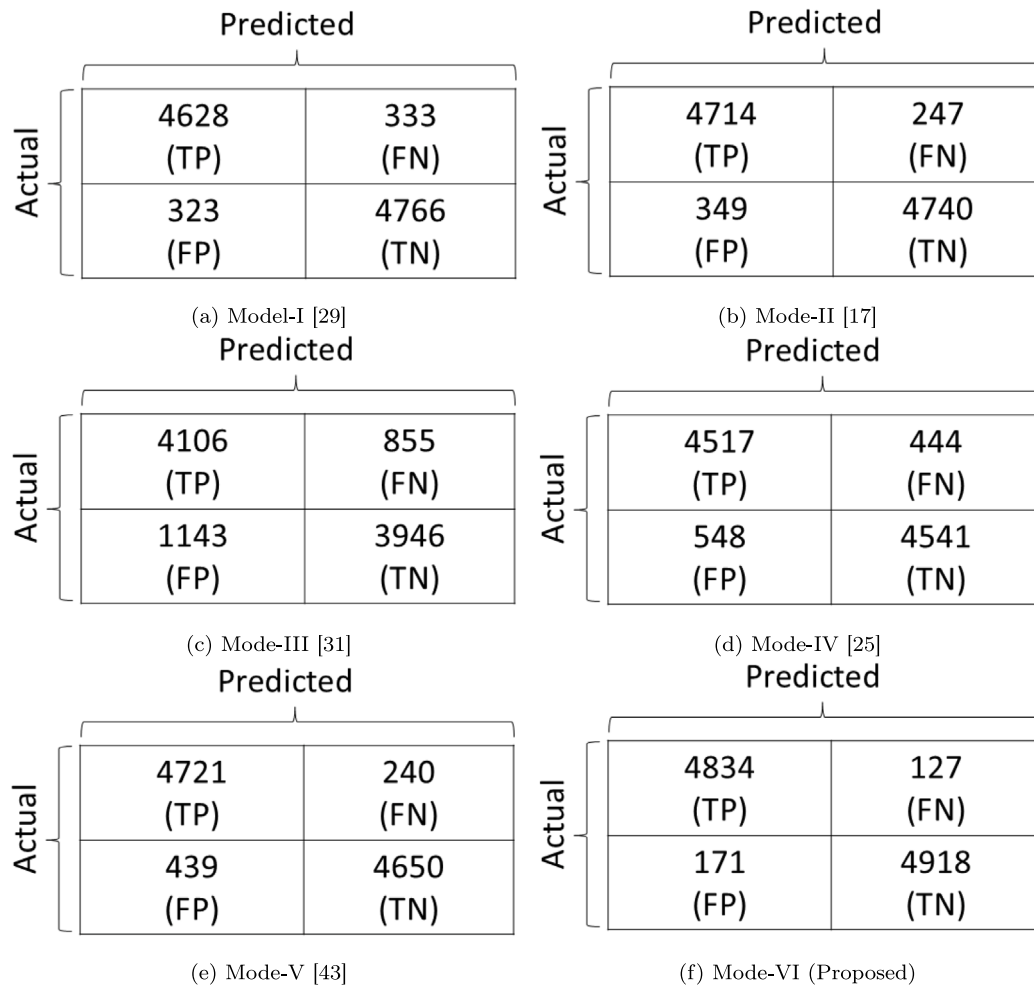


Fig. 6. Confusion matrices for the implemented models listed in Table 10.

cases, the proposed method achieved the highest accuracy measure 97.03%. Further validation has been done by comparing the result with five state-of-the-art ensemble models from other studies. Our proposed model outperformed all the existing models used for the medical diagnosis data. The research can be used in other healthcare domains with numerical features like diabetes, breast cancer, routine blood tests, thyroid, heart disease, and many more. The proposed data analytics pipeline will offer a complete dataset with less noise and a more reliable diagnosis result.

Though the study was done on the healthcare dataset, the implementation methodology can be used in other sectors such as finance, e-commerce, telecommunication, education, and the environment to deal with incomplete and noisy data. Using the hyperparameters suggested for the models, the ensemble part can also be replicated. To address the limitation of the research, it was conducted using numerical data. Further research needs to be done on unstructured data in healthcare like images, texts, or audio.

In the future, we plan to replicate the research in other domains as well to identify the robustness of the method. The data analytics pipeline suggested through the research may work differently in some cases. There is a research scope to answer why data preprocessing methods are separate for different domains of data.

CRedit authorship contribution statement

Lomat Haider Chowdhury: Writing – original draft, Visualization, Project administration, Methodology, Investigation.

Shaira Tabassum: Methodology, Formal analysis, Data curation, Conceptualization. **Swakkhar Shatabda:** Writing – review & editing, Supervision, Formal analysis. **Ashir Ahmed:** Writing – review & editing, Supervision, Resources, Conceptualization.

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.

Acknowledgments

The authors would like to acknowledge the Social Tech Lab of the Department of Advanced Information Technology, Kyushu University for providing their Portable Health Cleaning (PHC) dataset and allowing the authors to access their resources.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- [1] Jiang Z, Yang S, Liu Z, Xu Y, Xiong Y, Qi S, et al. Coupling machine learning and weather forecast to predict farmland flood disaster: A case study in yangtze river basin. *Environ Model Softw* 2022;105436. <http://dx.doi.org/10.1016/j.envsoft.2022.105436>.

- [2] Ali MH, Sarker MJ, Rahman MS, Rabbi F, Hossen MS, Kabir MA. Design and development of a gps-guided spray machine for reducing pesticide use on agricultural land in bangladesh. In: 2022 IEEE 12th symposium on computer applications & industrial electronics. IEEE; 2022, p. 66–70. <http://dx.doi.org/10.1109/ISCAIE54458.2022.9794549>.
- [3] Chowdhury LH, Farin NJ, Islam S. A deep learning approach to analyze the stock market during covid-19. In: International joint conference on advances in computational intelligence. Springer; 2022, p. 643–53. http://dx.doi.org/10.1007/978-981-99-1435-7_52.
- [4] Mukta MSH, Islam S, Shatabda S, Ali ME, Zaman A. Predicting academic performance: Analysis of students' mental health condition from social media interactions. Behav Sci 2022;12(4):87. <http://dx.doi.org/10.3390/bs12040087>.
- [5] Adhani A, Anshori A, Mahardika A. Public attitudes towards the government's policy communication in preventing covid-19. J Aspikom 2022;7(1):61–70. <http://dx.doi.org/10.24329/aspiikom.v7i1.1058>.
- [6] Chavda VP, Vuppu S, Mishra T, Kamaraj S, Patel AB, Sharma N, et al. Recent review of covid-19 management: Diagnosis, treatment and vaccination. Pharmacol Rep 2022;74(6):1120–48. <http://dx.doi.org/10.1007/s43440-022-00425-5>.
- [7] Doshi P, Godlee F, Abbasi K. Covid-19 vaccines and treatments: we must have raw data, now. BMJ 2022;376. <http://dx.doi.org/10.1136/bmj.o102>.
- [8] Al-Ali D, Elshafeey A, Mushannen M, Kawas H, Shafiq A, Mhaimeed N, et al. Cardiovascular and haematological events post covid-19 vaccination: A systematic review. J Cell Mol Med 2022;26(3):636–53. <http://dx.doi.org/10.1111/jcmm.17137>.
- [9] Mazza MG, Palladini M, Poletti S, Benedetti F. Post-covid-19 depressive symptoms: epidemiology, pathophysiology, and pharmacological treatment. CNS Drugs 2022;36(7):681–702. <http://dx.doi.org/10.1007/s40263-022-00931-3>, Springer.
- [10] Fadida BS, Sourani DL, Sira LB, Miller E, Bashat DB, Joskowicz L. Partial annotations for the segmentation of large structures with low annotation cost. In: Workshop on medical image learning with limited and noisy data. Springer; 2022, p. 13–22. http://dx.doi.org/10.1007/978-3-031-16760-7_2.
- [11] Karthika K, Lakshmi GJ. Prediction and classification of aerosol deposition in lung using ct scan images. In: AIP conference proceedings, 2519, AIP Publishing LLC; 2022, p. 253–65. http://dx.doi.org/10.1007/978-3-031-23683-9_18.
- [12] Sharma S, Guleria K, Tiwari S, Kumar S. A deep learning based convolutional neural network model with vgg16 feature extractor for the detection of alzheimer disease using mri scans. Measurement: Sens 2022;100506. <http://dx.doi.org/10.1016/j.measen.2022.100506>, Elsevier.
- [13] Vankar P. Forecasted U.S. national health expenditure as percentage of gdp. 2024. <https://www.statista.com/statistics/934320/us-health-expenditure-as-percent-of-gdp-forecast/>. [Accessed on 24 December 2024].
- [14] Department SR. Per capita health expenditure in selected countries. 2024. <https://www.statista.com/statistics/236541/per-capita-health-expenditure-by-country/>. [Accessed on 20 December 2024].
- [15] Al Gahtani A, El-Sayed H. Building a novel ensemble learning-based prediction framework for diagnosis of coronary heart disease. Int J Intell Syst Appl Eng 2022;10(3):265–73. <https://par.nsf.gov/biblio/10498670>.
- [16] Rai HM, Chatterjee K, Dashkevych S. The prediction of cardiac abnormality and enhancement in minority class accuracy with imbalanced ecg signals using modified deep neural network models. Comput Biol Med 2022;106142. <http://dx.doi.org/10.1016/j.cmpbiomed.2022.106142>, Elsevier.
- [17] AlJame M, Ahmad I, Imtiaz A, Mohammed A. Ensemble learning model for diagnosing covid-19 from routine blood tests. Inform Med Unlocked 2020;21:100449. <http://dx.doi.org/10.1016/j.imu.2020.100449>.
- [18] Callahan A, Shah NH. Machine learning in healthcare. In: Key advances in clinical informatics. Elsevier; 2017, p. 279–91. <http://dx.doi.org/10.1016/B978-0-12-809523-2.00019-4>.
- [19] Shailaja K, Seetharamulu B, Jabbar M. Machine learning in healthcare: A review. In: 2018 second international conference on electronics, communication and aerospace technology. IEEE; 2018, p. 910–4. <http://dx.doi.org/10.1109/ICECA.2018.8474918>.
- [20] Kiourtis A, Mavrogiorgou A, Menychtas A, Maglogiannis I, Kyriazis D. Structurally mapping healthcare data to hl7 fhir through ontology alignment. J Med Syst 2019;43:1–13. <http://dx.doi.org/10.1007/s10916-019-1183-y>.
- [21] Mavrogiorgou A, Kiourtis A, Touloupou M, Kapassa E, Kyriazis D. Internet of medical things (iomt): acquiring and transforming data into hl7 fhir through 5 g network slicing. Emerg Sci J 2019;3(2):64–77. <http://dx.doi.org/10.28991/esj-2019-01170>.
- [22] Saripalle R, Runyan C, Russell M. Using hl7 fhir to achieve interoperability in patient health record. J Biomed Inform 2019;94:103188. <http://dx.doi.org/10.1016/j.jbi.2019.103188>, Elsevier.
- [23] Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. 2018, p. 559–60. <http://dx.doi.org/10.1145/3233547.3233667>.
- [24] Al-Jamimi HA. Synergistic feature engineering and ensemble learning for early chronic disease prediction. IEEE Access 2024;12:62215–33. <http://dx.doi.org/10.1109/ACCESS.2024.3395512>.
- [25] Raza K. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In: U-healthcare monitoring systems. Elsevier; 2019, p. 179–96. <http://dx.doi.org/10.1016/B978-0-12-815370-3.00008-6>.
- [26] Kiourtis A, Nifakos S, Mavrogiorgou A, Kyriazis D. Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to hl7 fhir through ontology matching. Int J Med Inform 2019;132:104002. <http://dx.doi.org/10.1016/j.ijmedinf.2019.104002>.
- [27] Kouremenou E, Kiourtis A, Kyriazis D. A data modeling process for achieving interoperability. In: International conference on e-health and bioengineering. Springer; 2023, p. 711–9. http://dx.doi.org/10.1007/978-3-031-62502-2_80.
- [28] De Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Heal Technol 2022;12(2):255–72. <http://dx.doi.org/10.1007/s12553-022-00639-w>, Springer.
- [29] Verma A, Mehta S. A comparative study of ensemble learning methods for classification in bioinformatics. In: 2017 7th international conference on cloud computing, data science & engineering-confluence. IEEE; 2017, p. 155–8. <http://dx.doi.org/10.1109/CONFLUENCE.2017.7943141>.
- [30] Misra P, Yadav AS. Impact of preprocessing methods on healthcare predictions. In: Proceedings of 2nd international conference on advanced computing and software engineering. 2019. <http://dx.doi.org/10.2139/ssrn.3349586>.
- [31] Alim MA, Habib S, Farooq Y, Rafay A. Robust heart disease prediction: a novel approach based on significant feature and ensemble learning model. In: 2020 3rd international conference on computing, mathematics and engineering technologies. IEEE; 2020, p. 1–5. <http://dx.doi.org/10.1109/ICoMET48670.2020.9074135>.
- [32] Saihood Q, Sonuç E. A practical framework for early detection of diabetes using ensemble machine learning models. Turk J Electr Eng Comput Sci 2023;31(4):722–38. <http://dx.doi.org/10.55730/1300-0632.4013>.
- [33] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19(6):1236–46. <http://dx.doi.org/10.1093/bib/bbx044>, Oxford University Press.
- [34] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015;349(6245):255–60. <http://dx.doi.org/10.1126/science.aaa8415>, American Association for the Advancement of Science.
- [35] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44. <http://dx.doi.org/10.1038/nature14539>, Nature Publishing Group.
- [36] Esteve A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nature Med 2019;25(1):24–9. <http://dx.doi.org/10.1038/s41591-018-0316-z>.
- [37] Azad MA, Islam S, Farid DM, Shatabda S. Layered ensemble learning for effective binary classification. In: Emerging technologies in data mining and information security: proceedings of IEMIS 2020, vol. 3, Springer; 2021, p. 1–9. http://dx.doi.org/10.1007/978-981-15-9774-9_1.
- [38] Mavrogiorgos K, Kiourtis A, Mavrogiorgou A, Kleftakis S, Kyriazis D. A multi-layer approach for data cleaning in the healthcare domain. In: Proceedings of the 2022 8th international conference on computing and data engineering. 2022, p. 22–8. <http://dx.doi.org/10.1145/3512850.3512856>.
- [39] Data4u E. Diagnosis of COVID-19 and its clinical spectrum — kaggle.com. <https://www.kaggle.com/datasets/einsteindata4u/covid19>.
- [40] Tabassum S, Abedin N, Maruf RI, Ahmed MT, Ahmed A. Improving health status prediction by applying appropriate missing value imputation technique. In: 2022 IEEE 4th global conference on life sciences and technologies. IEEE; 2022, p. 345–8. <http://dx.doi.org/10.1109/LifeTech53646.2022.9754794>.
- [41] Islam R, Kikuchi K, Sato Y, Izukura R, Nishikitani M, Jahan N, et al. Portable health clinic system for maternal and child health care in covid-19 pandemic situation. Stud Health Technol Inform 2022;213–6. <http://dx.doi.org/10.3233/SHTI220700>.
- [42] Islam R, Yokota F, Nishikitani M, Kikuchi K, Sato Y, Izukura R, et al. Portable health clinic covid-19 system for remote patient follow-up ensuring clinical safety. Comput Methods Programs Biomed Updat 2022;100061. <http://dx.doi.org/10.1016/j.cmpbup.2022.100061>.
- [43] Yekkala I, Dixit S, Jabbar M. Prediction of heart disease using ensemble learning and particle swarm optimization. In: 2017 international conference on smart technologies for smart nation. IEEE; 2017, p. 691–8. <http://dx.doi.org/10.1109/SmartTechCon.2017.8358460>.