Original Research Paper

A 2-Tier Stacking Ensemble Classifier for Disease Classification

K. Sukanya Varshini and R. Annie Uthra

Department of Computational Intelligence, Faculty of Engineering and Technology,
SRM Institute of Science and Technology, Kattankulathur, Chennai, India

Corresponding Author:
K. Sukanya Varshini
Department of Computational
Intelligence, Faculty of
Engineering and Technology
SRM Institute of Science and
Technology, Kattankulathur,
Chennai, India

Email: sukanya.varshini@gmail.com

Abstract: Diarrhea, dysentery, and dehydration are top of the list of high mortality-causing diseases in children under the age group of five. Despite the tremendous growth of machine learning in the field of medical research, still, some areas remain untouched, especially, the pediatric department. Since medical data is used in the proposed work, even a slight increase in the accuracy of the model will be of great importance. In this study, we propose a 2-tier stacking ensemble method for disease identification. Initially, the data is pre-processed and sent to train the tier-1 machine learning models. Based on the majority voting the metadata is selected and sent to the next tier and finally, at the meta-classifier level disease classification happens. Performance metrics like the accuracy, precision, recall, F1-score and mean absolute error of the individual machine learning algorithms were analyzed and were used to compare with the proposed stacking ensemble method. The results proved that the 2-tier stacking model proposed in the work shows an accuracy of 95.41%, a precision of 94.47%, a recall value of 92.78%, and an F1-score of 93.50%. The proposed model achieved a high accuracy value when compared to the other machine learning models.

Keywords: Machine Learning, Disease Identification, Ensemble Stacking Model, Dysentery, Diarrhea, Dehydration, Pediatric

Introduction

As per the UNICEF report, more than a million children lose their lives in India before celebrating their fifth birthday (WHO) it is important to note that most diseases are easily preventable and treatable. Though humankind has been showing constant development in the reduction of the under-five death rate, the majority of the deaths happen in Sub-Saharan African countries. According to a recent survey in 2019 most of the under-five deaths occurred in the following five countries: Nigeria, India, Pakistan, the Democratic Republic of Congo, and Ethiopia, of which Nigeria and India account for most of the deaths.

Among the total 5.2 million deaths in 2019, children of age 1-11 months accounted for 1.5 million deaths and 1.3 million deaths were recorded for children of age 1-4 years and newborns accounted for 2.4 million deaths. The remaining deaths were from other age groups of 5-9 years. The primary reason for mortality among children in the under-five age group in the year 2016 includes acute respiratory infections, intrapartum-related complications, congenital anomalies, and diarrhea. Figure 1 portrays the comparative study of under-five deaths in Southeast Asian countries for the years 2015 and 2019 respectively,

in which India ranks top in the number of deaths of children under age five.

The above-mentioned scenarios have been considered by the government of India and rigorous action is taken by the Ministry of Health and family welfare under the national health mission. To reduce the mortality rate of neonates and children the document known as the Integrated Management of Neonatal and Childhood Illness (IMNCI) (NRHM) was drafted.

Based on the statistics mentioned above, it is very clear that the mortality rate of pediatric patients is very high and this can be reduced by the early detection and treatment of the disease. These statistical reports served as a motivation to pursue research on pediatric data in order to help in solving the problem to a large amount. The main contribution of the work is that the mortality rate of the children will shrink to a greater extent as the disease will be identified at an earlier stage with a better accuracy rate. The proposed work considers the dataset based on the diseases like dehydration, dysentery, and diarrhea that cause a high death rate in pediatric patients under the age of five.

The objective of the work is to develop a 2-tier ensemble method for disease identification in pediatric patients with better accuracy. The real-time dataset is

collected from Primary Health Center (PHC) and more data points are populated based on the Generative Adversarial Network (GAN) method. The disease classification in tier-1 is carried out using Naive Bayes (NB), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and the Extreme Gradient Boosting (XGBoost) and algorithms and the metadata are generated through voting technique. In tier-2, the Gaussian model is combined with the metadata from tier-1 to produce the second level of metadata. Finally, Random Forest (RF) is used as the meta-classifier for disease identification.

The main contributions of the paper are summarized as follows:

1. Data is collected from PHC and pre-processed for pediatric diseases based on the IMNCI (Integrated Management of Neonatal and Childhood Illness) document
2. Fivefold cross-validation is used to generate the metadata by feeding into the tier-1 stacking ensemble method comprising the Logistic Regression (LR), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and the Extreme Gradient Boosting algorithms (XGBoost)
3. The metadata generated from the tier-1 stacking method is combined with the Gaussian algorithm to produce the next set of metadata. This forms the tier 2 stacking model
4. The final set of metadata is fed into the meta-classifier for the final disease identification

Even though many researchers have carried out multiple works in disease identification using various machine-learning methods, still some areas remain untouched. One such area is the pediatric domain. Since there is not much work carried out on pediatric diseases, the research work related to other adult diseases is being addressed. Heart disease is the most analyzed disease using the stacking models by most of the researchers and the same is considered here in the literature survey. Many methods exist that help in the early detection of the disease and also estimate the readmission possibility of the patient. Based on the method of random forest, (Hsich *et al.*, 2011) have developed Random Survival Forest (RSF). The Gradient Boosting Machine (GBM) (Song *et al.*, 2016) is mainly dependent on the Gradient Boosting Decision Tree (GBDT) method which is used in various fields. GBDT always performs well on numerous classification benchmark datasets. XGBoost (eXtreme Gradient Boosting) is a version of GBM and it serves as an extension of GBM the same is proposed by Morise *et al.* (1992).

Medical experts are still struggling with the diagnosis of heart disease. Even though many countries are showing much research progress in the medical domain still the issue do exist in many poor countries (Polat *et al.*, 2007). Due to this high mortality rate has been recorded, so prediction and diagnosis of heart disease should be possible everywhere.
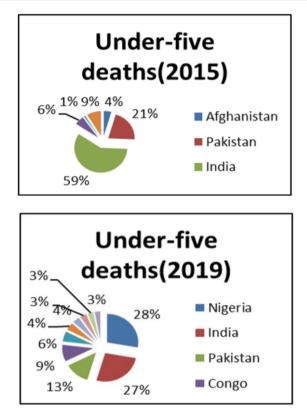

Fig. 1: Top South East Asian countries with the highest percentage of mortality for children under five years in 2015 and 2017; (Source: WHO)

Many traditional methods do exist in the ML field. Fuzzy Logic in addition to the ML algorithms also exists to enhance the performance metrics of the model. Fuzzy Logic is another important domain that is used more widely along with ML (Verma and Pal, 2020; ALzubi *et al.*, 2019).

Haq *et al.* (2019) proposed a stacking model for Heart Disease (HD) prediction. The sequential forward selection method was used for the feature selection and the Machine Learning (ML) algorithms used in the stacking model included Linear Regression (LR), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). The author concluded that the error rate of the stacking ensemble model decreased by 6.3% when compared to RF which was the best base learner. Balaji (2021) used the LR, support vector machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), RF, XGBoost, and Adaptive Boosting (AdaBoost) as the tier-1 ML algorithms in the model proposed by the researchers and showed an accuracy of 98%.

Aloyuni (2021) worked on a hybrid ML model for the identification of Coronary Artery Disease (CAD). The proposed model uses the base algorithms which include the DT, RF, and XGBoost. Cleveland dataset from the UCI repository is used. The proposed model showed an

accuracy of 98.34% in CAD prediction. Heikal and Eldawlatly (2020) proposed a stacking ensemble model for mortality rate prediction using the MIMIC dataset. The ML algorithms used here include the DT, Multilayer Perceptron (MLP), KNN, and LR. The method achieved an accuracy of 94.4%.

Based on the RF model, Chellamani et al. (2022) have developed Random Survival Forest (RSF). In this study, the Cleveland heart disease dataset has been used for HD detection. The RSF model proved that the algorithm had better performance metrics when compared to other ML algorithms as it showed a better accuracy of 94.3% when compared to the individual ML algorithms. The Gradient Boosting Machine (GBM) (Latha and Jeeva, 2019) is mainly dependent on the Gradient Boosting Decision Tree (GBDT) method which is used in various fields. GBDT always performs well on numerous classification benchmark datasets. XGBoost (eXtreme Gradient Boosting) is a version of GBM and it serves as an extension of GBM the same is proposed by Latha and Jeeva, (2020) in 2016 for the identification of coronary disease.

Yadav et al. (2020) the authors proposed a heart disease identification model in Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers. The model achieved an accuracy rate of 80.41% for the presence and absence of heart disease. A combination of ANN, hybrid ANN, and the fuzzy network was used to propose a model for the identification of heart disease and the model achieved an accuracy rate of 87.4%. In Kuruvilla and Balaji (2021) researchers developed a model for the identification of heart disease using ML models. The classifiers Naive Bayes (NB), Artificial Neural Network (ANN), and Decision Tree (DT) accomplished an accuracy of 86.12, 88.12 and 80.4% respectively. In Burse et al. (2019) the author proposed a model for heart disease identification using the ANN model and achieved an accuracy of 88.89%. Latha and Jeeva (2019) proposed a stacking ensemble model using RF, GBDT, and XGBoost as base learners for heart disease identification. HD diagnosis system (Yadav and Pal, 2020) was proposed based on the stacking ensemble method. The algorithms used as base learners include the KNN, LR, and NB. The final results of the ensemble were compared with the ML algorithms such as the bagging and boosting where the stacking model showed higher accuracy of 75.1%. An ensemble model of Fuzzy and ANN was proposed by the researchers for the HD diagnosis and achieved an accuracy rate of 91.10% (Tama et al., 2020). Djerioui et al. (2020) projected the HD model for the identification of heart disease based on LSTM technique and is compared with the MLP, where LSTM achieved a high accuracy rate of 89.1% Shorewala (2021). Medina-Quero et al. (2018) worked on the Cardio Vascular Disease (CVD) identification using Fuzzy Logic and RNN (El Sheikh et al., 2021; Haq et al., 2019; Kim et al., 2021).

The reason for selecting 2-tier-stacking ensemble models is that from the literature survey done it is clear that the performance rate of the ensemble model seems to be better than the individual machine learning algorithms when used individually. The higher the number of ML algorithms stacked together higher the accuracy rate of the model. Since the data that is analyzed falls in the medical domain even a slight increase in the accuracy of the model is very significant.

Proposed Method

The proposed method aims to identify the diseases such as dehydration, diarrhea, and dysentery using the stacking ensemble method from the medical dataset. The proposed method involves three different steps in the disease identification process, namely the data generation followed by the tier-1 stacking model, the tier 2 stacking model, and finally the meta-classifier. Initially, the data is generated using the random function based on the IMNCI document drafted by World Health Organization (WHO). The generated data corresponds to the pediatric patients who fall under the age group of 2 months to 60 months (i.e., 5 years). The generated data is preprocessed to check for the presence of NULL values, duplicate values, missing values, and outliers. The preprocessed input is fed into the novel two-tier stacking ensemble method. In the first tier, the data are cross-validated using five-folds and are fed into the tier-1 stacking ensemble method built using the logistic regression, random forest, Gradient Boosting Decision Tree (GBDT), and the Extreme Gradient Boosting algorithms (XGBoost). The metadata obtained from the tier-1 stacking model is combined with the Gaussian algorithm to form the tier 2 stacking ensemble model. Finally, the metadata obtained from the tier 2 stacking model is then fed into the Meta classifier for the final prediction. The performance metrics of the stacking ensemble method were analyzed by comparing the individual machine learning algorithms and Gaussian algorithms with the stacking ensemble method.

Dataset Description, Generation and Pre-Processing

Dataset Description

The dataset used in the proposed work is the real-time dataset collected from the Primary Health Centre (PHC) in and around the areas of Rani pet and Kanchipuram districts. The real-time dataset was collected in reference to the IMNCI (NRHM) document. IMNCI document acts as a guide to help the Primary Health Centre (PHC) to treat patients when doctors are not available around the clock. The staff nurse takes the responsibility of the PHC providing an initial level of treatment. This document helps them a lot in assisting the staff nurse by giving all the required details like the stage of the disease, which can be classified based on the symptoms of the disease,

medications for the various diseases along with the dosage. IMNCI document was prepared with the aim to reduce the mortality of children under five. The goal of the IMNCI document was to lower the death rate for children under the age of five. The IMNCI document is an adaptation of the WHO-framed Integrated Management of Childhood Illness (IMCI), where the Neonatal was incorporated in the Indian version. Infants under two months old are considered neonates. Kids between the ages of 2 months and 5 years belong to the children category. The sample IMNCI document is shown in Fig. 2.

Dataset Generation Using GAN Method

As machine learning algorithms are used in the proposed work, the real-time dataset that was collected was insufficient for the model to perform well. So, the synthetic data generation method using the Generative Adversarial Network (GAN) was adopted. GAN uses two models called the generator and discriminator. The work of the generator is to generate fake data very similar to real data in such a way that the discriminator does not identify the difference between fake and real data. After a certain time, the discriminator will not be able to differentiate between real and fake data. Here, the generative model captures the data distribution and is trained to try to increase the likelihood that the Discriminator will make a mistake. On the other hand, the Discriminator is based on a model that calculates the likelihood that the sample it received came from the training data and not the generator. The Discriminator is seeking to reduce its reward $V(D, G)$ in the minimax game that the GANs are designed as, while the Generator is trying to maximize its loss by minimizing the Discriminator's reward. Figure 3 gives the simple architecture of GAN. The following Eq. (1) can be used to mathematically describe it:
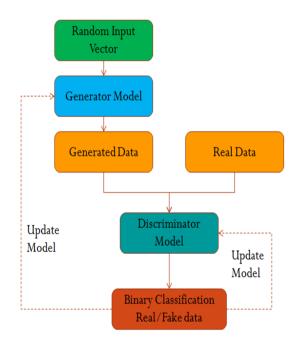
$$V(D,G) = E_{x \sim pdata(x)}\left[\log D(x)\right] + E_{z \sim pdata\ (z)}\left[\log\left(1 - D(G(Z))\right)\right] \quad (1)$$

where, G and D are the generator and discriminator. $G(Z)$ is the generator network, and $D(x)$ is the discriminator network.

The count of the real-time dataset was 150 and after the synthetic generation of data, a total of 5000 data were produced. The proposed method considers the features like temperature, age, lethargy, restlessness, sunken eyes, drinking poorly, drinking eagerly, skin pinch slow, skin pinch very slow, blood in stools, duration of diarrhea, and dehydration to identify the diseases like dehydration, dysentery, diarrhea. After the synthetic generation of data, the data is checked for the existence of duplicate values, null values, and missing values. The data got after the pre-processing step is fed into tier-1 of the stacking model.

Fig. 2: IMNCI document

Fig. 3: A simple GAN working model

Two-Tier Stacking Ensemble Method

Let $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ be the dataset, where, $x_i \in X$, $y_i \in Y = \{c_1, c_2, \ldots, c_l\}$. N is defined as a complete set of instances $c_1, c_2, \ldots, c_l$ are defined as the labels of the class. The input space X contains a number of features and hence its elements are represented as N-tuple of 'd' dimensions. Figures 4-5 shows the architecture of the stacking ensemble model.
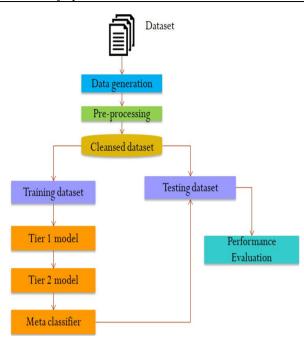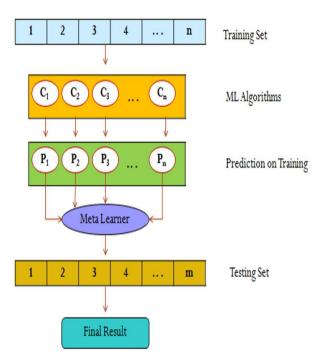
Fig. 4: A general block diagram

Fig. 5: A simple stacking ensemble model

$$X = \left\{ \left(x_{11}, x_{12}, \ldots, x_{1d}\right), \left(x_{21}, x_{22}, \ldots, x_{2d}\right), \ldots, \left(x_{N1}, x_{N2}, \ldots, x_{Nd}\right)\right\} \quad (2)$$

Here d is defined as the total feature set. The below representation denotes the dataset:

$$X = \left\{ \left(\left(x_{11}, x_{12}, \ldots, x_{1d}\right), y_1\right), \left(\left(x_{21}, x_{22}, \ldots, x_{2d}\right), y_2\right), \ldots, \left(\left(x_{N1}, x_{N2}, \ldots, x_{Nd}\right), y_N\right)\right\} \quad (3)$$

Fragment the dataset D as S_1 and S_2. Where S_1 is the training set and S_2 is the testing set, they are divided into 70 and 30% respectively:

$$D = S_1 \cup S_2 \quad (4)$$

For tier-1 the classifiers are modeled in the following way.

Let $W_C = \{W_{C1}, W_{C2}, \cdots, W_{Cm}\}$ represent the classifiers in tier-1 of the model.

Let T_1, T_2, $\cdots$, and T_5 represent the set that has been divided from set S_1.

Fold 1: The training set $T = T_1 \cup T_2 \cup \ldots T_5$

The testing set $V = T_5$

The training set T is used further to train the ML algorithms in tier-1:

$$W_{CK} : Xg \to Y$$
$$W_{CK} : X_g \to \left\{C_1, C_2, \ldots, C_l\right\} \quad (5)$$

where:

$$k = 1, 2, \ldots, m \text{ and } X_g \in T \quad (6)$$

The subsequent fold, T_5 is considered for the testing purpose whereas the other sets are used for training purposes. The steps are again done further till every fragmented dataset is worked as a testing set. Thus, the prediction result of every dataset is received.

Let $W_L = \{W_{L1}, W_{L2}, \cdots, W_{Lm}\}$ represent the labels of the tier-1 classifier:

$$i.e.\ W_{LK} = W_{CK}(X_g) \quad (7)$$

where, $k = 1, 2, \ldots, m$.

The mathematical model for the ensemble tier is given as follows.

Let $E_C = \{E_{C1}, E_{C2}, \cdots, E_{Cm}\}$ are the combination factors. The labels produced by the ML algorithms in tier-1 will be mapped into a single class:

$$E_{CK} : W_{LK} \to Y \quad (8)$$

$$E_{CK} : W_{LK} \to \left\{C_1, C_2, \ldots, C_l\right\} \quad (9)$$

$$E_{CK} = Aggregation\ of\ (W_{LK}) \quad (10)$$

$$E_{CK} = Aggregation\ of\ (W_{CK}(X_g)) \quad (11)$$

where, $k = 1, 2, \ldots, m$.

Let $E_L = \{Y_{ec1}, Y_{ec2}, \cdots, Y_{ecm}\}$ represent the class labels which was identified by tier-2. The same has been represented below:

$$Y_{eck} = E_{CK}(W_{LK}) \qquad (12)$$

$$Y_{eck} = E_{CK}(W_{CK}(X_g)) \qquad (13)$$

where, $k = 1, 2, …, m$.

The meta classifier considers the features: $C_F = \{C_{F1}, F_2, F_3\}$, ensemble predictions E_L and the label Y as input.

Ensemble methods are meta-algorithms that integrate many machine learning algorithms into a single model. Many techniques do exist in the method of combining ML algorithms to produce a new ensemble algorithm to provide better prediction results when compared to the individual algorithms. The proposed system uses the stacking ensemble technique where multiple algorithms, typically different, machine learning algorithms are piled up together to produce a more powerful method for more accurate results through the process of meta-learning. In the proposed work, two-tier stacking ensemble methods are used for more accurate disease identification of pediatric diseases. Two-tier stacking ensemble methods comprise three stages. The first stage involves the process of combining base learners namely Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GBDT), and Extreme Gradient Boosting (XGBoost) to generate metadata. Then these metadata and Gaussian algorithms are combined to form another set of metadata. Finally, this metadata is fed into the meta-classifier for the accurate identification of diseases. The overall block diagram is given in Fig. 6.

The steps involved in the 2-tier stacking ensemble model are listed below in 3 steps

The Ensemble Emodel is Set Up

1. The base learners of the tier-1 stacking model are specified
2. The base learners of the tier-2 stacking model are specified
3. The meta classifier is specified

Tier-1 Base Learners are Trained

a. Every base larner in tier-1 is trained with the dataset.
b. Five-fold cross-validation is performed on every base learner in tier-1 and the predictions are retrieved from every base learner used in tier-1 (Note that the same number of folds should be implemented). Let these predicted values be $p_1, …, p_K$
c. The five-fold cross-validated data from every base learner in tier-1 are fused together to create a new matrix. The final result is combined with the original vector and this results in the data which is received from tier-1:

$$\left[p_1 \right] … \left[p_K \right][y] \rightarrow [Z][y] \qquad (14)$$

where, Z is the representation of the $N \times L$ matrix.

Train the Ensemble Tire-2

a. Train the Gaussian algorithm with the same training set and the same number of k-fold cross-validation as used in tier-1
b. Combine the level one data and the output of the Gaussian algorithm to form "level two" data
c. The meta classifier is trained with the data from tier 2 ($y_1 = f(Z_1)$)

Predict New Data

a. To generate ensemble predictions, first generate predictions from the base learners
b. Combine the predictions from the level 1 model with the Gaussian algorithm prediction
c. The final results are used in the meta-classifier for the final prediction results

The 2-tier stacking ensemble model involves four steps namely setting the ensemble method, training base learners in tier-1, training base learners in tier 2, and finally, the classification using the meta-classifier. The first step involves the selection of the base Learners (L) involved in tier-1 and the base learners (M) involved in tier 2 of the stacking model along with the meta-classifier which is used at the final stage of the stacking model.

The next step is the training of the L base learners in the tier-1 stacking model with the dataset set that has been generated initially. The data has been trained with fivefold cross-validation on every ML algorithm that has been used in tier-1. The predicted values from every algorithm have been collected and given as input in the next level. every base learner used in tier-1. An important point to be noted is that the same number of five-folds are used for every algorithm in both tiers. Let the predicted values be $p_1, …, p_k$. The N cross-validated predicted values from each of the L algorithms can be combined to form a new $N \times L$ matrix (represented by Z). This matrix, along with the original response vector (y), is called the "level-one" data, where N is the number of rows in the training set.

The next step is the training of the M base learners in the tier-2 stacking model with the dataset set that has been generated at the early stage and used in the tier-1 stacking model. The data will be trained with the five-fold cross-validation. Train the Gaussian algorithm with the same training set and the same number of k-fold cross-validation as used in tier-1. Combine the level one data and the output of the Gaussian algorithm to form "level two" data. Now the training dataset for the meta-classifier is ready. Train the meta-learning algorithm on the level-two data ($y_1 = f(Z_1)$).
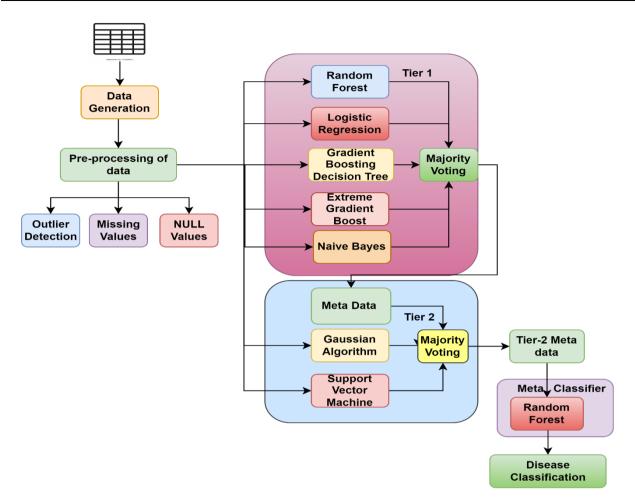
Fig. 6: Block diagram highlighting the overall working of the proposed work

The final classification is made based on the meta-classifier that has been selected. To generate ensemble predictions, first, generate predictions from the base learners. Combine the predictions from the level 1 model with the Gaussian algorithm prediction. Feed those predictions into the Meta learner to generate the final prediction.

Tier-1 Stacking Ensemble Method

The tier-1 stacking model involves the usage of base learners including the Logistic Regression (LR), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Naïve Bayes (NB), and the Extreme Gradient Boosting algorithm (XGBoost). These base learners all together are trained and five-fold cross-validation is used to generate the metadata. All the base learners are trained with the same set of data and the same number of folds in cross-validation. Once the training for different base learners is done, these four base learners generate four prediction lists. The generated four prediction lists are now combined to produce a new list that serves as a training dataset for the base learner in the Tier 2 stacking model.

The various functions involved among the various base learners of the tier-1 stacking model are discussed below.

Tier-2 Stacking Ensemble Method

The metadata that has been generated from the tier-1 model is combined with the Gaussian algorithm in the second tier of the stacking ensemble model. The same set of training datasets is used to train the Gaussian algorithm and the metadata obtained from the tier-1 model. This tier 2 stacking model produces a second set of metadata. This metadata is finally fed to the Meta classifier which is the third and final step in the two-tier stacking model. The Gaussian distribution used in tier 2 of the stacking ensemble model is given below.

The data set which was generated previously was shuffled randomly and has been split into 5 folds. For every fold which was separated, one set of the fold will be treated as test data, and the remaining folds are taken for the training set. The whole process as mentioned in the algorithm is repeated for the Gaussian algorithm. The algorithm is mainly comprised of two loops. The first loop

is used in the process of consisting of the algorithm to be trained and the loop present after it will be focusing on the five-fold cross-validation process to produce both the training and the test data. Training data is also split into five folds. One of the folds of the training data is taken to be the validation test and the remaining folds that are present will be considered as the training data set. The training dataset is used in the process of training the base-level model. Then in the next step, the data enters into the base-level model which is present to generate the test 1. As the loop is found to repeat 5 times, it is considered to train the base level model at the sum of 5 folds and the average of the test data along with the metadata from tier-1 is generated. The output is calculated by taking the union of both the training set and the testing set of the base model.

Algorithm 1: 2-Tier stacking ensemble model

Input: Training dataset: $X_i = \{(x_1, c_1), (x_2, c_2), …, (x_i, c_i)\}$
    Tier-1 classifiers: $L_1, L_2, …, L_n$
    Tier-2 classifiers: *Guassian G1*
    Meta classifier: Random Forest *RF*
Output: Trained ensemble classifier *M*
BEGIN
Step 1: Train the tier-1 classifier by applying classifiers L_n to the dataset X_i
        for $i = 1, 2, … k$ do
            $T_{1i} =$ mode $\{L_j(X_i)\}$ for $j = 1, 2, … n$
        end for
Step 2: Pass the metadata obtained from step 1 D_1 to train the Tier-2 classifier
        for $i = 1, 2, … K$ do
    % use T_{1i} to classify the training example X_i %
            $Z_i = T_{1i}(X_i)$
            $\hat{D} = \{Z_i, C_i\}$
            where $T_{2i} = G(T_{1i})$
        end for
Step 3: Train a metalevel classifier *RF*
        $\hat{RF} = RF \{\hat{D}\}$
        vote $\{T_{1i}, T_{2i}, \hat{D}\}$
        return $\hat{M}$
end

Meta Classifier

The metadata obtained from tier 2 of the stacking model is finally fed to the Meta classifier which is the third and final step in the two-tier stacking model. Here Random Forest is used as a Meta learner so that it is trained with the new metadata that was generated and tested using the test data to produce the final results. The two-tier stacking ensemble algorithm has been explained in detail in Algorithm 1. It is to be noted that the main purpose of using the fivefold cross-validation while training the base learners is to make sure that the data of the new training set is distributed as consistently as

possible. Otherwise, there is a possibility that the Meta learner will overfit the new dataset and makes a biased prediction in the Meta learner stage. Meta learner is used in the stacking ensemble method to find the optimal combination of the algorithms. The main objective of the ensembling technique is to combine two or more diverse algorithms into a single stacking method to produce better results.

Linear Regression (LR), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and the Extreme Gradient Boosting (XGBoost) algorithms were selected as base learners for the tier-1 stacking model. Among these base learners selected, LR is a classical linear model whereas the remaining algorithms are non-linear models. The models were selected heterogeneously as base learners in the first stage to increase the diversity and the robustness of the proposed stacking model, which will have the capability to learn the training data and interpret the data from various aspects. In tier 2 Gaussian algorithm is combined with the metadata obtained from the tier-1 level and another set of metadata is produced. At last Random Forest is selected as a Meta learner in the next stage of the model, such that it can become accustomed to the advantages of the base learners and make better results.

The performance metrics of the Gaussian algorithm and the individual algorithms i.e., Linear Regression (LR), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and the Extreme Gradient Boosting (XGBoost) used as base learners used in tier-1 were also analyzed to compare it with the proposed 2-tier stacking method.

Materials and Methods

The proposed work is completed related to Machine Learning models and no materials are associated with the work proposed. And related to methods, the proposed 2-tier stacking ensemble model is already discussed in the section Two-Tier Stacking ensemble model.

Results and Discussions

Evaluation Metrics

The performance metrics for the multiple criteria, like the accuracy, precision, recall, and F1-score of the individual machine learning algorithms were used to compare with the proposed 2-tier-stacking ensemble model. The work is found to be executed in the Jupiter notebook in Python. The performance metrics were calculated based on the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The formulas are given below:

$$Accuracy = \frac{TP}{TP + FP + TN + FN} \quad (15)$$

$$Precision = \frac{TP}{TP + FP} \qquad (16)$$

$$Recall = \frac{TP}{TP + FN} \qquad (17)$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (18)$$

Results of Data Preprocessing

The proposed method involves four different steps in the disease identification process. Initially, the data is generated using the random function concerning the IMNCI document drafted by World Health Organization (WHO). The data was generated based on the range values in the IMNCI document. The generated data is then preprocessed to check for the presence of NULL values, duplicate values, missing values, data imbalance, and outliers. The same has been given in Fig. 7. The generated data is cross-validated using five-folds and is then fed into the Stacking ensemble method built using the Logistic Regression, Random Forest, Gradient Boosting Decision Tree (GBDT), and the Extreme Gradient Boosting (XGBoost). Finally, the metadata is obtained which is then fed into the Meta classifier for the final prediction. Then the performance metrics of the stacking ensemble method were analyzed. The performance of individual machine learning algorithms and Gaussian algorithms were used to compare with the stacking ensemble method. The performance metrics analyzed were accuracy, precision, recall, and F1-score.

Pair Plot and KDE Plot

Based on the data generated based on the IMNCI document, the relationship between every variable can be plotted. A pair plot plots a pair-wise relationship in a dataset. The pair plot function creates a grid of axes such that each variable in data will be shared on the y-axis across a single row and on the x-axis across a single column. Datasets under real-time study contain many variables.

Kernel Density Estimation (KDE) is a method to generate a smoothing curve for a given set of data. This method can be used to visualize the shape of certain data as a replacement for the histogram. The KDE algorithm takes a parameter, bandwidth that affects how "smooth" the resulting curve is. Use the control below to modify the bandwidth and notice how the estimate changes.

The KDE is calculated by weighting the distances of all the data points seen for each location on the blue line. If the points are near, the estimate is higher, indicating the probability of seeing a point at that location. Figure 8 gives the pair plot of temperature and age. Figures 9-10 give the KDE plot for various diseases.
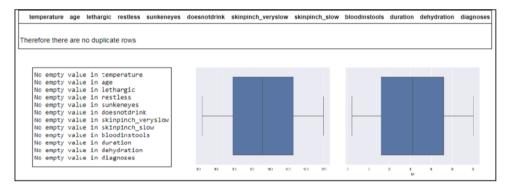

Fig. 7: Data preprocessing steps performed after data generation. Duplicate values were checked for every feature and no duplicate values exist. Null values are also checked along with the outliers and the results are shown in the above diagrams
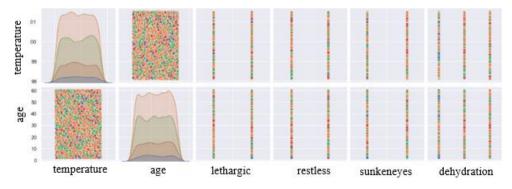

Fig. 8: The pair plots of temperature and age concerning various features have been shown in the figure
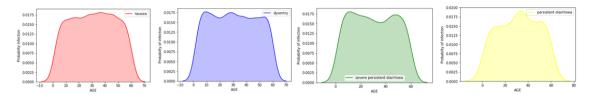
Fig. 9: The KDE plot is given for the various disease based on the key parameter age to understand the density distribution of the data
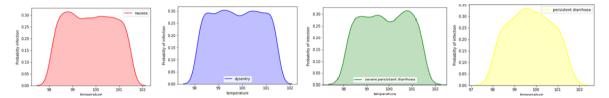

Fig. 10: The KDE plot is given for the various disease based on the key parameter temperature to understand the density distribution of the data

Once the preprocessing of the data was done, the data generated were split into 70% for training and 30% for testing. Later the training data is split into 5 sets to perform the k-fold cross-validation on each of the base learners and the predicted values were saved and these predictions were fed into the Meta-learning algorithm to generate the final predictions. The performance measures like the accuracy, precision, recall, F1-score, sensitivity, specificity, balanced F measure AUC ROC curve and mean absolute error have been tested with the individual machine learning algorithms along with the Gaussian algorithm to compare with the results of the proposed stacking method.

Results of Algorithms Used in 2-Tier Stacking Method

In the 2-tier stacking method, multiple algorithms are used as base learners in both tier-1 and tier-2 levels. In the tier-1 stacking model, the base learners used include the Logistic Regression (LR), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost). These algorithms are trained with the training data with five-fold cross-validation. These tier-1 base learners produce the metadata. Then this metadata is combined with the data obtained from the Gaussian algorithm which is the base learner of tier-2. Then the final metadata is obtained which is then fed into the meta-classifier to produce the final result. The same dataset and the five cross-validation scheme is used by every algorithm in the proposed model. The performance metrics like the accuracy, precision, recall, F1-score, sensitivity, specificity, balanced F measure AUC ROC curve and mean absolute error have been calculated and the results have been discussed below. The confusion matrix and the ROC curves for the individual algorithms have also been analyzed.

Results of Algorithms in Tier-1 Stacking Model

Results of Logistic Regression Classifier

The logistic regression algorithm being the first base learner in the tier-1 stacking model is trained with the dataset that has been generated initially and the testing data is used to check the various performance measures. The data has been divided into 70:30 ratios for training and testing respectively. When analyzed with the various performance metrics it measured an accuracy of 95%, with sensitivity and specificity values recording 73% and 70% respectively. The mean absolute error measured around 0.078. Balanced F-measure and AUC ROC curve are recorded as 71 and 85% respectively. Figures 11-12 show the confusion matrix and ROC curves of the logistic regression algorithm.

Results of Random Forest Classifier

The Random Forest algorithm one of the base classifiers has been trained with the same dataset that has been used by the linear regression. It is made sure that the same dataset and cross-validation are used by every algorithm in tier-1 and tier-2 levels. Random forest classifier measured the accuracy of 96% and the value of sensitivity and specificity recording 75 and 72% respectively. 73 and 87% value has been recorded for the balanced F measure and AUC ROC curve respectively. Figures 13-14 show the output obtained by the random forest classifier.

Results of Extreme Gradient Boosting Algorithm

XGBoost algorithm has been trained with the dataset that has been generated initially. The XGBoost algorithm has been tested with various performance measures and it measured an accuracy of 95%, with sensitivity and specificity values recording 73 and 70% respectively. 71 and 85% value has been recorded for the balanced F measure and AUC ROC curve respectively. The accuracy calculated by this algorithm showed lower accuracy than the other base learners in the tier-1 algorithm. Figures 15-16 show the output obtained by the XGBoost algorithm.
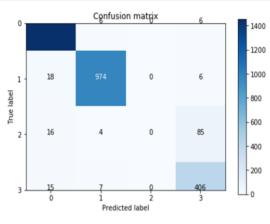
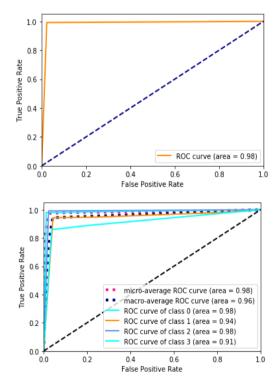**Fig. 11:** Confusion matrix of logistic regression classifier

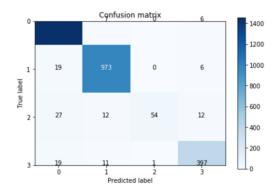**Fig. 12:** ROC curves of linear regression classifier
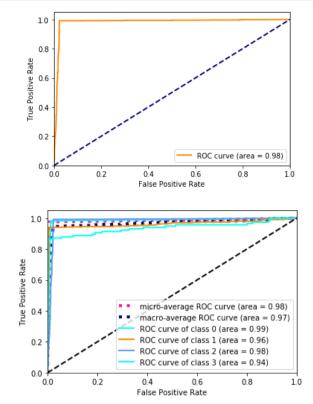

Fig. 13: Confusion matrix of random forest classifier
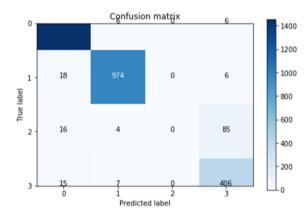

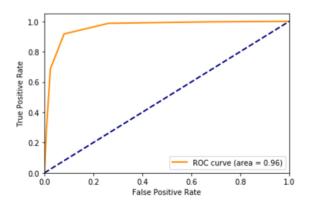
Fig. 14: ROC Curves of random forest classifier

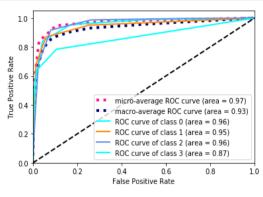**Fig. 15:** Confusion matrix of XGBoost classifier

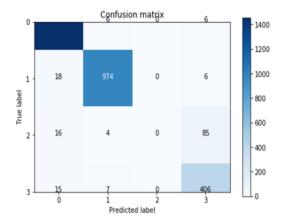**Fig. 16:** ROC Curves of XGBoost classifier

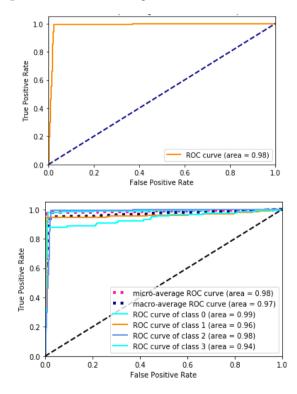**Fig. 17:** Confusion matrix of gaussian classifier

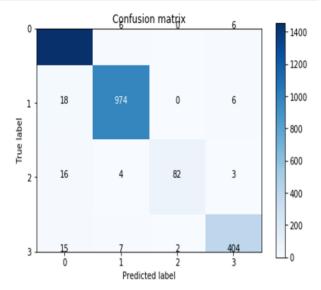**Fig. 18:** ROC curves of Gaussian classifier

Fig. 19: Confusion matrix of the 2-tier stacking ensemble model

Results of Algorithm in Tier 2 Stacking Model

Results of the Gaussian Algorithm

The Gaussian algorithm acts as the base learner in the tier 2 stacking model and is trained with the dataset that has been generated initially and the testing data is used to test the performance measures. The data has been divided into 70:30 ratios for training and testing respectively. The Gaussian algorithm was used to test the efficiency of the algorithm for the generated data. The algorithm has measured the accuracy of 96%, with sensitivity and specificity values recording 85 and 96% respectively. 89 and 91% value has been recorded for balanced F-measure and AUC ROC curve respectively. Figures 17-18 show the output obtained.

Results Obtained by the Meta Classifier

The proposed 2-tier-stacking ensemble algorithm finally involves the meta-classifier. The meta-classifier used in the proposed method is the model that finally outputs a set of metadata which is the combination of the result obtained from the Gaussian algorithm and the metadata from the tier-1 stacking model. This final set of metadata from the tier 2 level is the training dataset for the meta-classifier. Finally, the testing dataset is used to test the performance of the meta-classifier which is also the final prediction result of the proposed 2-tier stacking model. The proposed method measured the accuracy of 95%, with sensitivity and specificity values recording 73 and 70% respectively. 71 and 85% value has been recorded for the balanced F measure and AUC ROC curve respectively. Figures 19-20 show the output obtained by the proposed method.
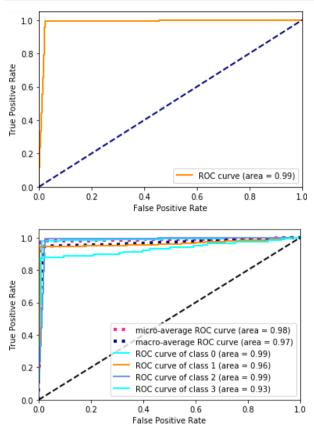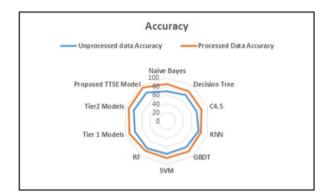
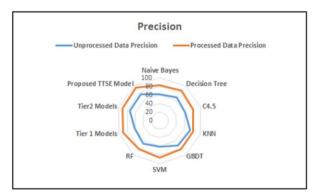**Fig. 20:** ROC Curves of the 2-tier stacking ensemble model

Fig. 21: Performance Metrics values obtained for individual ML Models before and after processing data

Results of the Proposed 2-Tier Stacking Model

The performances of the base learners and our proposed stacking model are shown in Figure 21. For base learners, the tree-based models have relatively better performances compared with the Gaussian algorithm and Linear Regression and the Random Forest has the best performance over the other base learners used in both tier-1 and tier-2 stacking models. From the analysis of the individual ML algorithms, it is concluded that the 2-tier stacking algorithm showed an increase in the accuracy rate when compared with the individual ML algorithms. As the dataset analyzed in the work includes the medical dataset, which is highly sensitive and even a small improvement in the accuracy rate matters, so using a 2-tier stacking model is highly recommended.

The proposed method is finally compared with the Cleveland dataset. Cleveland dataset consists of the data for heart disease prediction. Table 1 gives the accuracy achieved by various researchers using the Cleveland dataset. The limitations, advantages, and accuracy achieved by the various researchers are given and lastly, our method is related and compared with the existing works.

The stacking model proposed in the work showed promising results when compared to the ML algorithms in tier-1. In particular, among base learners, Random Forest showed higher result values when compared with the other base learners. The Gaussian algorithm also showed higher values when compared with the individual base learners. But in the end, the proposed algorithm showed better results when compared with the other base learners and the Gaussian algorithm showed an accuracy of 97%.

Table 1: Comparison of accuracy achieved by various researchers using the cleveland dataset

Reference	Technique used for heart disease detection	Limitations	Advantages	Accuracy (%)
Morise *et al*. (1992)	LR	Low accuracy	Low computation time	77.00
Polat *et al*. (2007)	ANN	More time for execution	Accuracy is high	84.50
ALzubi *et al*. (2019)	Ensemble method	Computationally complex	High accuracy	89.01
Medina-Quero *et al*. (2018)	ANN, Fuzzy	More execution time required	Accuracy is high	91.91
Kuruvilla and Balaji (2021)	MLP and SVM	Computationally complex	High accuracy	80.41
Burse *et al*. (2019)	Hybrid Neural, fuzzy neural, ANN	Hybrid neural performance low	ANN achieved high accuracy	87.40
Latha and Jeeva (2019)	NB, ANN, DT	DT showed low accuracy	NB showed high accuracy	86.12
Yadav and Pal (2020)	ANN	More time for execution	Accuracy is high	84.89
El Sheikh *et al*. (2021)	ANN, fuzzy	More execution time required	Accuracy is high	91.10
	Proposed method	Computationally complex	Accuracy is high	94.60

Table 2: Values of performance metrics after feature selection through t-test

Classification model	Accuracy	Specificity	Recall	F1-score	Precision
K-nearest neighbour	85.76	82.79	80.90	87.60	84.25
Naive Bayes	83.01	80.90	78.80	85.80	82.76
SVM	95.05	92.80	93.20	92.65	92.01

Figure 21 gives the overview of the results obtained from the individual base learners, Gaussian, and Stacking methods. Comparison has also been made with every algorithm to prove the efficiency of the proposed method.

The result obtained gives us the picture that the ensembling method performs well and has higher prediction accuracy. Still, in a field like the medical field, even a minute increase in prediction accuracy pays high importance and is significant. High-volume datasets with an increased number of cross-validation folds can be considered to get a higher accuracy rate in the performance of the algorithms.

A t-test is a type of inferential statistic used to determine the significant difference between the means of two groups, which may be related to certain features. A t-test is used as a hypothesis testing tool, which allows testing an assumption applicable to a population. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance. A t-test allows us to compare the two data sets' average values and determine if they came from the same population.

In the proposed model, T-test is used to select the best features that are more suitable for the disease classification. The best features were ranked according to the significance level. These features were classified by SVM using a 10-fold cross-validation rule. 11956 iterations were performed and the process was repeated until the threshold number of clusters was reached. Each test result was evaluated as sensitivity, specificity, and accuracy. The performance measures were measured and the highest classification accuracy was found to be 95.1459% after selecting the features. At the same time, the data was classified without any preprocessing. The classification performance metrics including sensitivity, specificity, accuracy, and recall are calculated in Table 2.

After the disease is identified using the 2-tier stacking ensemble model, the Medicine Recommendation System (MRS) is proposed. This MRS system prescribes the medicine based on the stage of the disease classified. MRS gets the details of the patient including name, age, temperature, and the symptoms of the patient through the web application. The inputs received by the web application are the symptoms that have been extracted from the IMNCI document. Based on the details entered, the disease and the stage of the disease are classified. Medicine prescriptions along with follow-up details are recommended for the patient. Figures 22-23 show the medicine prescribed by MRS for severe dehydration and dysentery. Similarly, the classification of another disease can also be done.

The complete automation of the MRS can be used by the staff nurses in Primary Health Centers where doctors are not available around the clock. This system will recommend the medication and the follow-up details for the patient based on the symptoms that have been observed. This system prescribes medicine based on the stage of the disease that has been classified.

Fig. 22: Web application showing the medication for severe dehydration

Fig. 23: Web application showing the medication for dysentery

Conclusion

In the proposed work 2-tier stacking model is proposed comprising the Linear Regression (LR), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost) as the base learners in tier-1 of the stacking model. Gaussian algorithm along with the tier-1 metadata works in tier-2 of the stacking model. Then the metadata from tier-2 is fed into the random forest classifier which acts as the meta-classifier. Data used in the proposed work has been generated based on the IMNCI document. Performance metrics like the accuracy, precision, recall, F1-score, sensitivity, specificity, balanced F measure, AUC ROC curve and mean absolute error of the individual base learners in both the tier of the stacking model is analyzed and used to compare with the proposed 2-tier stacking ensemble method. The results that were obtained proved that the 2-tier stacking model proposed in the work showed promising performance as stacking of many best-performing algorithms in 2-tiers showed a higher accuracy rate than the algorithms that were used individually in the identification of diseases. The work will be carried forward in the future by analyzing the real-time dataset and also concentration will be given to the neonatal data and prescription of medicine according to the disease identified.

Acknowledgment

Funding Information

Author's Contributions

K. Sukanya Varshini: Contributed to the introduction, and literature survey and implemented the results for the research article.

R. Annie Uthra: Contributed invalidated the results and workflow of the research work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Aloyuni, S. A. (2021). A Systematic Review on Machine Learning and Deep Learning Based Predictive Models for Health Informatics. *Journal of Pharmaceutical Research International*, *33*(47B), 183-194.
https://doi.org/10.9734/jpri/2021/v33i47B33112

ALzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., & Thaventhiran, C. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. *Applied Soft Computing*, *80*, 579-591.
https://doi.org/10.1016/j.asoc.2019.04.031

Balaji, T. (2021). An insight on machine learning algorithms for predicting heart diseases. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(10), 5867-5877.
https://doi.org/10.17762/turcomat.v12i10.5405

Burse, K., Kirar, V. P. S., Burse, A., & Burse, R. (2019). Various preprocessing methods for neural network-based heart disease prediction. In *Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS-2018* (pp. 55-65). Springer Singapore.
https://doi.org/10.1007/978-981-13-2414-7_6

Chellamani, M., Murugesan, S., & Raju, N. (2022, October). Heart disease prediction using Boosting Algorithms: Performance Analysis and Comparison. In *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)* (pp. 1-6). IEEE.
https://doi.org/10.1109/MysuruCon55714.2022.9972536

Djerioui, M., Brik, Y., Ladjal, M., & Attallah, B. (2020, September). Heart Disease prediction using MLP and LSTM models. In *2020 International Conference on Electrical Engineering (ICEE)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICEE49691.2020.9249935

El Sheikh, A., Mahmoud, N., & Keshk, A. E. (2021). Heart Disease Classification Based on Hybrid Ensemble Stacking Technique. *IJCI. International Journal of Computers and Information*, 8(2), 1-8. https://ijci.journals.ekb.eg/article_207732_327e1ee8ddec8a289150c0ddedd1104d.pdf

Haq, A. U., LI, J., Khan, J., Memon, M. H., Parveen, S., Raji, M. F., ... & Monday, H. N. (2019, December). Identifying the predictive capability of machine learning classifiers for designing heart disease detection system. In *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing* (pp. 130-138). IEEE. https://doi.org/10.1109/ICCWAMTIP47768.2019.9067519

Heikal, M., & Eldawlatly, S. (2020, December). An Ensemble Classification Technique of Neurodegenerative Diseases from Gait Analysis. In *2020 15th International Conference on Computer Engineering and Systems (ICCES)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICCES51560.2020.9334609

Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., & Lauer, M. S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes*, 4(1), 39-45. https://doi.org/10.1161/CIRCOUTCOMES.110.939371

Kim, C., You, S. C., Reps, J. M., Cheong, J. Y., & Park, R. W. (2021). Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. *Journal of the American Medical Informatics Association*, 28(6), 1098-1107. https://doi.org/10.1093/jamia/ocaa277

Kuruvilla, A. M., & Balaji, N. V. (2021, February). Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach. In *IOP conference series: Materials Science and Engineering* (Vol. *1085*, No. 1, p. 012028). IOP Publishing. https://doi.org/10.1088/1757-899X/1085/1/012028

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, 100203. https://doi.org/10.1016/j.imu.2019.100203

Medina-Quero, J., Zhang, S., Nugent, C., & Espinilla, M. (2018). Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition. *Expert Systems with Applications*, 114, 441-453. https://doi.org/10.1016/j.eswa.2018.07.068

Morise, A. P., Detrano, R., Bobbio, M., & Diamond, G. A. (1992). Development and validation of a logistic regression-derived algorithm for estimating the incremental probability of coronary artery disease before and after exercise testing. *Journal of the American College of Cardiology*, 20(5), 1187-1196. https://doi.org/10.1016/0735-1097(92)90377-Y

NRHM. Operational Guidelines for Facility-Based Integrated Management of Neonatal and Childhood Illness (IMNCI). https://nhm.gov.in/images/pdf/programmes/child-health/guidelines/imnci_chart_booklet.pdf

Polat, K., Şahan, S., Kodaz, H., & Güneş, S. (2007). Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism. *Expert Systems with Applications*, 32(1), 172-183. https://doi.org/10.1016/j.eswa.2005.11.024

Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, 26, 100655. https://doi.org/10.1016/j.imu.2021.100655

Song, R., Chen, S., Deng, B., & Li, L. (2016). eXtreme gradient boosting for identifying individual users across different digital devices. In *Web-Age Information Management: 17th International Conference, WAIM 2016, Nanchang, China, June 3-5, 2016, Proceedings, Part I 17* (pp. 43-54). Springer International Publishing. https://doi.org/10.1007/978-3-319-39937-9_4

Tama, B. A., Im, S., & Lee, S. (2020). Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. *BioMed Research International*, *2020*. https://doi.org/10.1155/2020/9816142

Verma, A. K., & Pal, S. (2020). Prediction of skin disease with three different feature selection techniques using stacking ensemble method. *Applied Biochemistry and Biotechnology*, 191(2), 637-656. https://doi.org/10.1007/s12010-019-03222-8

WHO. The Global Health Observatory-Explore a world of health data. https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/child-mortality-and-causes-of-death

Yadav, D. C., & Pal, S. A. U. R. A. B. H. (2020). Prediction of heart disease using feature selection and random forest ensemble method. *International Journal of Pharmaceutical Research*, 12(4), 56-66. https://doi.org/10.31838/ijpr/2020.12.04.013

Yadav, S. S., Jadhav, S. M., Nagrale, S., & Patil, N. (2020, March). Application of machine learning for the detection of heart disease. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 165-172). IEEE. https://doi.org/10.1109/ICIMIA48430.2020.9074954