International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

An Approach to Prediction of Cardiovascular Diseases using Machine and Deep Learning Models

*1 Bhagyalaxmi Singirikonda, 2Dr.Muktevi Srivenkatesh

Submitted: 18/08/2022 **Accepted:** 26/11/2022

Abstract : In this article, we examined approximately 550 patient records in order to determine major risk variables that may be the root cause of cardiac issues. This study attempts to offer a piece of work that may be used as an instant step toward determining a probability assessment for the heart condition. The many risk factors are those that may be the major cause of the emergence of a cardiac condition. In this work, we examined several classification methods to diagnose the heart condition. The data was gathered from five separate Indian cities and also from people of all ages. We used the bidirectional LSTM model (BDLSTM), which was trained with experimental data from the source. The primary goal of this type of activity is to present a clear solution that will allow the patient to know statistically whether a heart problem is likely to occur. This answer is not a substitute for a healthcare professional, but rather a supplement to any doctor's diagnostic procedure. This look after transparency in the treatment of a physician and a patient. The model is evaluated based on both positives and negatives of false generated by model, and the best method for prediction is chosen. The results showed that two of these four algorithms were more accurate than 98.5% of the time.

Keywords: *Healthcare, feature selection, Machine Learning, cardiovascular diseases*

1. Introduction

WHO stated that people around 17.5 million had lead to death in 2019 by the CVDs , this accounts 30% of overall mortality. Cardiovascular diseases kills more people per year than any other cause. Coronary heart disease causes 7.4 million CVDs, whereas stroke, hypertension, cardiovascular disease, rheumatic heart disease, and heart failure cause 6.7 million. CVDs mostly affect poor and middle-income nations. By 2030, CVDs are seen to be main cause for mortality from the worlds below par nations. (The Top 10 Causes of Death, 2020). Simply put, machine learning is making the healthcare system smarter. Through its capacity to learn associations, this strong mini set of AI can be recognizable to several in use cases akin to speech recognition utilized by voice assistants and developing tailored on-line searching experiences. However, machine learning has shown potentially life-changing promise in healthcare, notably in medical diagnosis. There are various barriers to rapid machine learning incorporation in healthcare today (**Habehh et al, 2021; Cheddad & Abbas, 2020**). One of the most difficult difficulties in getting patients data to requisite the quality and size of ML working models. Furthermore, the issues arises with structure of data will often wanted by the substantial work to organize and clean in ML analysis. The healthcare business is progressively expanding the capabilities of machine learning and data science as they become more widely used. Because each patient produces high volumes of clinical information including results of X rays, immunizations, genomes, medication lists, other past medical history, and much more, to involve the data analysis will be the primary function. Also, a large data is created, stored and taken using many different domains that include healthcare, applications through mobile etc,. (**Daoudy et al.2019**)

Storing, processing, displaying, and extracting information from such enormous and diverse data kinds has become difficult with insufficient state-of-the-art technology tools. Exploring methods to successfully get relevant information for diverse sorts of consumers be a critical technical problems of large information analysis. At the moment, numerous types data sources of healthcare type are been gathered in every setting of non clinical and clinical, with a record that includes the medical history of a patient being an essential information source in analytics of healthcare. As the result, creating daata system distributed to handle huge amounts of data has three major challenges: For starters, it is challenging to gather data from dispersed areas owing to the diverse and massive amount of data. Considering the storage will be most difficult issue modifying diverse, huge amounts of data in the model. A System which contains big data must give guaranteeing performance as well as store data. The last problem is connected to BDA, namely taking enormous data from all the time it can be real or virtual for modelling, predictions, optimizing, visualization etc,.

ML provides a more comprehensive approach, since a model may be created that enables an intelligent solution to learn new rules. The techniques utilized in ML are mostly statistical in nature. DL is a subgroup of ML that deals with the usage of neural networks, with the word "deep" referring to the number of layers in the network. It is difficult to apply ML to this

1*Research Scholar, Department of Computer Science, GITAM School of Science, GITAM University, Visakhapatnam , Andhra Pradesh, India

2Associate Professor, Department of Computer Science , GITAM School of Science, GITAM Deemed to be University, Visakhapatnam, Andhra Pradesh, India

1 bhagyalaxmigit@gmail.com

massive data stream because typical ML algorithms isn't designed to take huge volumes and vary in pace. In healthcare, ML is utilized to tackle a variety of challenges (**Tekkeşin and Ahmet, 2019; Sidey et al. 2019; Puaschunder and Julia, 2020**) heart disease is individualized, so that the rigidity of heart diseases changes from person to person (**Keto et al. 2016**). As a result, developing an ML type, trains its own dataset, so the information of patient's may be solved in the prediction and is dependent on data provided , so it doesn't apply to other person. Diabetes of Type 2 can be a condition that can be avoided as it can be maintained by lifestyle and health weight (**Olokoba et al. 2019**). Using machine learning algorithms (MLAs) and deep learning can help predict heart attacks more accurately (**Ahishakiye et al. 2012**).

In order to look at this particular study, we will focus on the following main goals:

1. To look at various ML algorithms, finding needed one that can predict the heart condition with the most accuracy and the fewest errors.
2. Determine the major risk factors or variables that contribute to cardiac problems and heart disease.
3. To monitor the heart's condition and predict CVD.

The suggested architecture is made up of five major components: Most of these models serve as the foundation for our ideas. To test the efficacy of the Machine learning algorithms on the datasets, thereby find the CVD risk elements, we used a six-stage technique. The following are the six stages that is loading the dataset followed bt preprocessing of data next by selecting the attributes followed by runing the ML models and adapting the assessment metrics and finally processing the classifier results by its performance.

2. Background And Related Work

CVD and DM that is diabetic mellitus and cardiovascular diseases now account for a considerable portion of the worldwide burden on health care also co-occur oftenly. Present techniques considerably fail in detecting patients who have both diabetes and CVD, resulting in a time loss for the therapy,also increase the complexity for chance of life.

Abdalrada et al. (2022) created and tested a ML model with 2 stage method to achieve the result for both CVD and DM. They employed a dataset from DiScRi with more than 199 variables taken by 2000 individuals. In the first step, they employed Evimp functions and also Logistic Regression integrated into a model to identify important shared risk variables for CVD and DM then can be a matrix of correlation to eliminate duplication. They employed algorithms like regression and classification is create a model in second step. Deep breathing heart rate change, Gender are not correct cosidering BP change, TCHDL ratio were all variables contains risk in common for CVD and DM variables in co-occurrence. This will give 94.09% precision whether DM and CVD will happen together. Its sensitivity is 93.5%, and its specificity is 95.8%.

Sung et al. (2022) compared the efficiency is based on survival analysis of both RNN and Cox regression to demonstrate the enhanced accuracy of deep learning. As the use of national health screening and insurance can choose 361,239 participants who had two or more health examination records stated from 2006 to 2002. (NHIS-HEALS). Average number of body-related screenings included in the study (from 2002 to 2013) was 2.9. From the NHIS-HEALS data, two CVD prediction models were created. The regression model of Cox had the greatest when time comes as an option to pick from the curve of AUC of 0.75 in men and 0.79 in females in the dataset available at the range of 2 years. The DL model had the been crucial when time is considered in the AUC curve 0.96 in males and 0.94 in females. In that order, LRP found that age was the most important factor in Cardio Vascular Disease are considered with SBP and DBP in order to get results.

Xie et al. (2021) introduced the DBSCAN method which is abbreviated as density-based spatial clustering of applications with noise. This is weighted strategy of learning to use dataset information of density to find accurate prediction for cardiovascular diseases. (CVDs). Random Forest (RF) technique is used to pick significant characteristics, then separates them into three different groups by dividing sample points using various uses through density by weight learning. As a result, ML models built with primary characteristics since the feature of weight can get dense data easity, also shows the boundaries for decision taking more effectively, also get superior activities. Cross-validation showed that ML models using weight learning can improve its accuracy by 3% which points to the Stroke data set.

Rustam et al. (2022) addressed this challenge by suggesting a unique application called convolution neural network which consists of a feature extraction. This model can extend characteristic set needed for linear models to train such as the logistic regression, gradient(stochastic) descent classifier and helps vector machines which compose the model for soft voting ensembling it. Tests are extensive as they are carried out for evaluating the functinality of various features in the training dataset. The performance of four distinct datasets is analysed, and the findings are compared to contemporary techniques utilized for CVDs. The suggested model outperforms the competition with 0.93 correctness and 0.92 for remaining scores. The output shows that technique that was suggested is better and that the ensemble model can be used with many different datasets.

Barbieri, et al. (2022) compared deep learning extensions of survival analysis models to The national health databases of administrators were predicted by the Cox proportional Hazards at CVD risk.The army creation of New Zealand people aged around 30 to 74 were engaged with health services provided publicly in the year 2012 using individual person linking of administrative databases. After taking out people who had CVD in the past, we used ML and proportional Hazards by Cox since figuring out how likely it was that CVD would happen within 5 years. The fraction variation explained, calibrating model and discriminating, the predictor variables are for hazard ratios which were used to include the results. Results: events that are occurred at CVD are 61 927 out of 2,164,872 individuals. Other found that predictors were consistent as existing risk of CVD factor information. Taking percentage as basis it is explained variation, discrimination and calibration since the deep learning surpassed models of Cox proportional Hazards.

majority papers either concentrate on a single healthcare data source on BOC called as Batch Oriented Computing. However, information about healthcare are diverse as constantly generate large amounts of data. As a result, integrated health information analysis, which includes machine learning power tools, realtime analysis, stream data collection etc,. which is requirement for constructing a system for commencing with dispersed streams that include health related data. Tools like real time analysing may minimize time taken for attendence physically and assist physicians in anticipating the possible sickness. Another key aspect of approach proposed will be as if the sickness of the patient isn’t good, The service is provided at he emergency immediately contacted by the technology(alert) to conduct steps at the event if emergency is called out.

3. Proposed Model

Figure 1 depicts the whole procedure of the suggested technique. Following data gathering and visualization, a two-stage feature selection procedure is carried out. The data is then standardized and divided into train and test sets to train, test the model. The next paragraph goes through the specifics of the suggested technique. To test the efficacy of the suggested approach,

A. Database:

The patient data set used in this article is a patient data set from various locations in India, having a total of 550 entries. Approximately 550 individuals were investigated throughout this investigation, and critical characteristics that may be the underlying cause of heart attacks were found. This research tries to offer a job that may be an impending step toward a possible cardiac problem. These particular risk factors are the primary causes of heart disease. To define the core problem in this research, many categorization systems were examined. The data was collected from five different cities and age groups in India. This dataset takes into account risk elements like family history, smoking, fasting glucose, diet, CABG, high serum etc,. The records are in binary, with 1 indicating yes and 0 indicating no. Along with this, there are three other data columns: age, sex, and location.

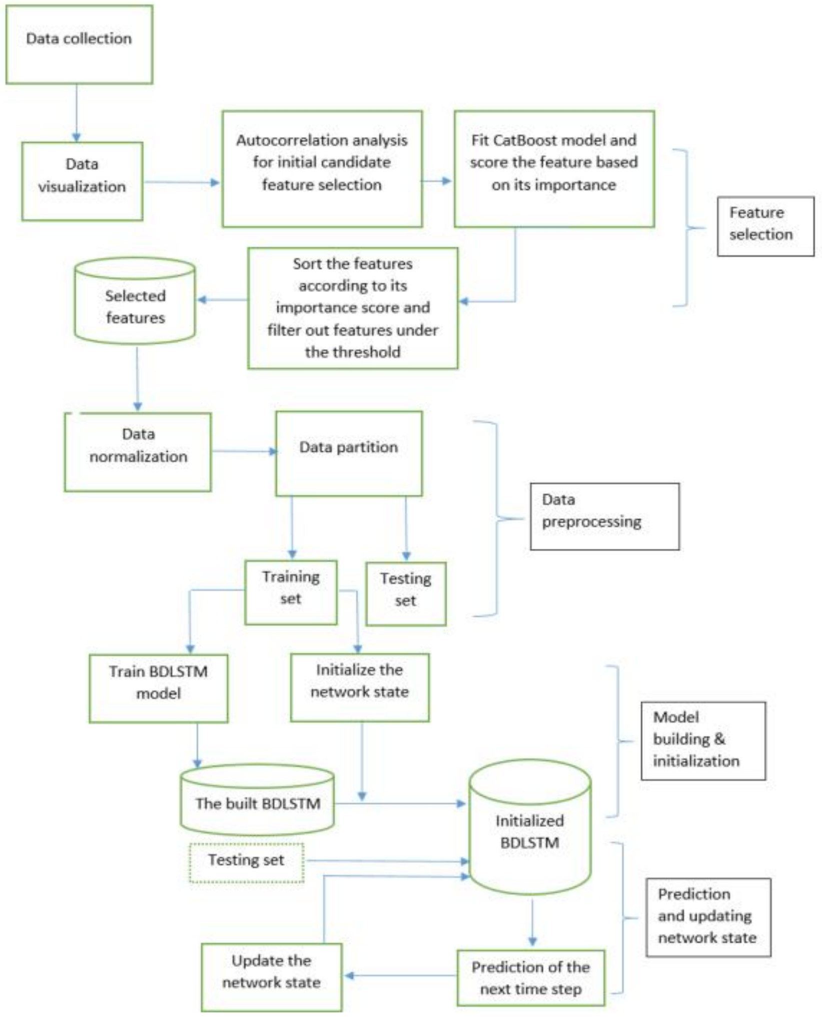
B. Methodologies for Data Preprocessing:

Various data preparation and imputation procedures must be used in order to prepare the dataset for input into the modeling and thus to continue the research job. Missing value imputation, outlier treatment, data scaling, polynomial feature creation, one-hot encoding, and other approaches are widespread. In the instance of data imputation, since the dataset is real-time and in the healthcare domain, the crucial element is to contact a domain expert rather than imputing it in terms of statistical measurements.

CatBoost as feature selection:

Traditional boosting methods, on the other hand, involve preprocessing procedures that transform into numeric representations by changing categorical input variables. To address this issue, Sratedy for effective boosting is most efficient is Cat Boost (**Dorogush et al. 2018**), is bought into action. CatBoost, precisely employs changed TBS algorithm called as target based algorithm.

Figure 1-The proposed scheme in the form of Block Diagram is shown in


Data collection
Data visualization
Autocorrelation analysis for initial candidate feature selection
Fit CatBoost model and score the feature based on its importance
Feature selection
Selected features
Sort the features according to its importance score and filter out features under the threshold
Data normalization
Data partition
Data preprocessing
Training set
Testing set
Train BDLSTM model
Initialize the network state
Model building & initialization
The built BDLSTM
Initialized BDLSTM
Testing set
Prediction and updating network state
Update the network state
Prediction of the next time step

Assume $D = \{(X_i, Y_i)\}_{i=1,\dots,n}$), where $X_i = (x_{i,1,\dots,}x_{i,m})$ Vector which is containing categorial and numeric information and has features of value m in it. The related label Yi 2 R. First, the dataset is permuted at random. Mean label value is then determined for samples provided in one permutation which belings to same category. We define$\sigma = \sigma_1, \ldots, \sigma_n$ represent the permutation. The permuted observation $x_{i,\,k}$ is then replaced with $x_{\sigma_j,k}$ and $x_{\sigma_p,k}$ is calculated as

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right]Y_{\sigma_j} + a.P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] + a} \qquad (1)$$

where $[xi_{,k} = xj_{,k}]$=11 if $xi_{,k} = xj_{,k}$ else 0 otherwise. '*P*' represents previous data, while 'a' represents associated mass. Advance denotes mean label value for regression. The a deductive likelihood of meeting label of positive value as classification. Priority helps which limit disturbances for smaller domains. A suggested technique trains on the whole dataset. On the other hand, by executing random permutations, it avoids the overfitting problem.

Furthermore, in CatBoost, a novel schema for computing leaf values while choosing a tree structure is given to solve the biased gradients difficulties in traditional boosting algorithms (**Li et al.2019; Friedman et al. 2002**). Suppose F^i signifies the constructed model $g^i(X_k, Y_k)$ and/or to maintain the gradient unbiased, M_k a separate model is trained and used for each sample X_k, Since it doesn't update a sample by using gradient estimation . M_k is used to determine the gradient on X_k, the generated tree is then evaluated based on the estimate. Algorithm 1 shows the detailed stages of the algorithm.

Algorithm 1: Estimation of Gradient by Cat Boost

```
1. Input: train data ${(X_k, Y_k)}_{k=1}^{n}$ since the random permutation is used, T that is the number of Trees,and function named as choice loss :$\varphi(y_j, a)$
2. Initialization: $M_i$ < - for $i = 1,\ldots,n$
3. Do for $iter = 1,\ldots,I$
Do for $i = 1,\ldots,n$
Do for $j = 1,\ldots,i-1$
        $g_i < -\frac{d}{da}\varphi(y_j, a)|_{a=M_i(X_j)}$
$M$ <- BuildOneTree $((X_i, g_i) for\ j = -1, \ldots. i, ..1)$
$M_i$ < - $M_i + M$
4. Output: $M_1(X_1)$ , $M_2(X_2)$ , …., $M_n(X_n)$;$M_1$, …., $M_n$ :
```

The first candidate qualities are sent into the Algorithm called as Cat Boost first to reduce characteristics that provide less helpful information for prediction. An equation is used to determine the relevance of each feature after model fitting (2).

$$Feat_{imp} = \sum_{trees,leaves}\left(\boldsymbol{v_1} - \frac{\boldsymbol{v_2 c_1 + v_1 c_1}}{\boldsymbol{c_2 + c_1}}\right)^{\mathbf{2}} .\boldsymbol{c_1} + \left(\boldsymbol{v_2} - \frac{\boldsymbol{v_2 c_2 + v_1 c_1}}{\boldsymbol{c_2 + c_1}}\right)^{\mathbf{2}} .\boldsymbol{c_2} \qquad \mathbf{(2)}$$

Ci where I belongs to 1and 2 signify leaf node samples and v_1, v_2 denote the leaf's formula value As the score for each feature is known precisely, it is put in order from most important to least important, a cutoff would be around 0.5 as a score used which filters out the lower importance features ones.

Data partitioning:

To start the modeling the whole data is divide into 2 sets which are train set which consists of 70% of data and the test set which consists of remaining 30 % of data to check the model is correctly trained for the unseen inputs given to it.

C. Building modeling:

Random Forest, Logistic Regression, KNN, and BDLSTM were used for analysis and prediction. In this scenario, the model was fed into the Grid Search and Random Search Algorithms with varied hyperparameter values. The optimum algorithm setup is attained through optimizing hyperparameters. The best model for each setting is then predicted, by comparing the findings the best model is taken out . Best model with hyper parameters is retained and used to evaluate real-time accuracy and performance on out-of-sample data. Analysis and prediction models help doctors diagnose more accurately. This supports physicians in making accurate diagnoses and reduces the diagnostic turnaround time.

KNN: Its is defined as K-Nearesr Nieghbours where we use this technique for instance-based supervise learning.This process takes the sluggish way of approaching a problem. As it doesn't generate models, Specific for collection of examples for labeled training. It evaluates lengths to generate models which is close to group of examples of labeled training. This also evaluates lengths for a given training instances to the whole training instances present. The test instance's lengths will be utilized to forecast label of a class. As all the class labels are added to the examples of K training that are closest to the test instance.

Random Forest (RF): Frim the subset of the predictors which are selected from a random node,It is divided into many by using best split. This process differs from that used in normal trees, in which every node gets divided among the characteristics present at the data set by using optimal split under consideration. Furthermore,Values which are projected by predicting it by merging the decision trees that have been created. It is an algorithm since it draws forecast its final taken by a larrge number separate models. These unique models might be of a similar or dissimilar nature. Because decision trees are employed in the random forest approach, the individual models are of the same type.

Logistic regression (LR): This model represents the probability for categorized tasks with two outcomes. It is a classification problem extension called as Linear Regression Model(LRM). On plus side, LRM provides probabilities in addition to classification. This gives it a big advantage over models that can only give the final classification. Considering that an instance has a 99% chance of belonging to a class as opposed to 51%, it has a major impact.

Algorithm:

i) From the training set, a bootstrap sample is chosen. In this bootstrap sample, an unpruned tree is generated.
iii) At each internal node, a number of nodes is chosen at random and the optimal split is calculated.
iv) The overall forecast is determined by the majority vote of all trees.

The concept of BDLSTM was developed from the notion of a bidirectional recurrent neural network (**Thireou et al. 2007; Graves et al. 2009**). If the training is given to both forward and backward for any two recurrent networks then it is a bidirectional recurrent neural network, which are coupled to the same output layer. In the same way, a BDLSTM has forward and backward LSTM layers that connect to the same output layer and process sequence data in both directions.

Figure 2 shows the deep learning model architecture. It is made up of four bidirectional LSTMs that are stacked on top of each other. B, L, and C are the batch size, the longest sequence length of the batch, and the feature size, respectively. The C represents the input data channel number. For every time, step *t*, given a mini-batch input $X_t = R^{n\times d}$ (number of examples: *n*, number of input features in each example: *d*), let the hidden layer activation function be Φ.

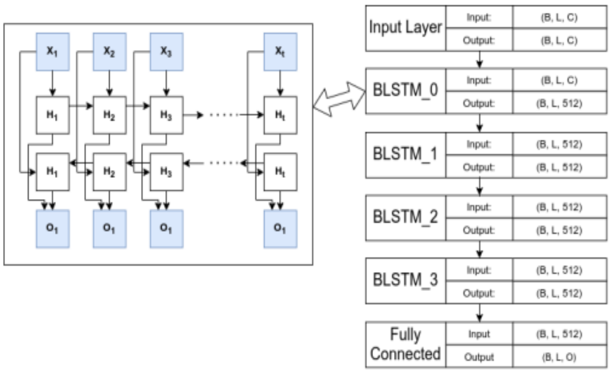
We suppose that the forward and backward hidden states for this time step in the one-layer bidirectional LSTM are $\vec{H}_t \in R^{n\times h}$ and $\overleftarrow{H}_t \in R^{n\times h}$, correspondingly, where *Hidden units is mentioned as h*. The following are forward and backward concealed state updates:

$$\vec{H}_t = \Phi\left(X_t W_{xh}^{(f)} + \vec{H}_{t-1} W_{xh}^{(f)} + b_h^{(f)}\right) \tag{3}$$

$$\overleftarrow{H}_t = \Phi\left(X_t W_{xh}^{(b)} + \vec{H}_{t-1} W_{xh}^{(b)} + b_h^{(b)}\right) \tag{4}$$

$W_{xh}^{(f)} \epsilon\, R^{d\times h}, W_{hh}^{(f)} \epsilon\, R^{h\times h}, W_{xh}^{(b)} \epsilon\, R^{d\times h}, W_{hh}^{(b)} \epsilon\, R^{h\times h}$ and biases $b_h^{(f)} \in R^{1\times h}$ *and* $b_h^{(b)} \in R^{1\times h}$ are all the model parameters.

Figure 2 *Architecture of bidirectional LSTM. The sub figure on the left represents a bidirectional LSTM cell.*



Following that, we concatenate the forward and backward hidden states $\vec{H}_t$ and $\overleftarrow{H}_t$ to generate the hidden state $H_t \in R^{n\times 2h}$ that will be passed into the next layer. Then repeat the calculation from equation 1. When we analyze dropout, the hidden state $H_t \in R^{n\times 2h}$ will be arbitrarily masked as zeros at every step during the training phase with a frequency of rate δ (in this study, the dropout rate is δ = 0.1. The dropout may simply assist in preventing overfitting. Finally, the output layer calculates $O_t \in R^{n\times q}$ (number of outputs: *q*):

$$O_t = activation\left(H_t W_{hq}\right) + b_q \tag{5}$$

The model parameters of the output layer are the weight matrix $W_{hq} \in R^{2h\times q}$ and the bias $b_q \in R^{1\times q}$.

After the features have been chosen, the BDLSTM state is set up by first trying to make predictions on the training phase. The one-time step prediction technique is then used at data points at the testing set where it checks for n time steps. To be more explicit, the initialised BDLSTM model predicts the first testing sample. Updation of the network state before predicting the modified model test cases for the time step, we are using prediction value. This method will be repeated in order to anticipate the remaining time steps.

The model parameters of the output layer are the weight matrix $W_{hq} \in R^{2h\times q}$ and the bias $b_q \in R^{1\times q}$. **D. Parameter tuning:**

n estimators and max features are the primary features to tune.The size that need to be examined as the splitting of a node happens of the random subsets features is specified by the max parameter. To increase the bias and also to reduction the variance is large where the max features value should be low. N estimators give the value of number of trees present in the random forest classifier .To achieve the value the number of trees should be more but it would be a long time process as it takes time to calculate. We will also discover that after a certain number of trees, the results will no longer improve considerably. numFeatures is the value of max features. The Gini index of the dataset may be determines the random features in numbers and divide them into a tree.

4. Results And Discussions

The empirical evaluation of the proposal is provided in this part, and the findings are reviewed. The experiment was carried out on a main dataset of 550 records obtained from various locations in India. Python 3 is used for the programming portion. This experiment was carried out in a PC running with 8GB of ram and runs on windows 10. Approximately 550 patients were investigated throughout this investigation, and significant indicators that may be the underlying reason for heart attacks were found. To define the core problem in this research, many classification systems were examined. The data was collected from five different cities and age groups in India. We use classification accuracy and confusion matrices to examine our findings. Confusion matrices are utilized to have a better understanding of a model's performance. It shows the outcome in a 2x2 matrix with cells denoting TP, TN, FP, and FN. The false negative would be diagnosing a patient as healthy when they are really unwell.

Metrics:

The important measures for measuring the efficiency of the models must then be defined. To represent the performance of classifiers diagnostic problem in the digital way of approaching we take four features (numbers) which can essentially create measurement for derivative measures.

FP (False Positive) – The healthy patients who are detected in the classifier,

TP (True Positive) – The unhealthy patients who are detected in the classifier,

FN (False Negative) – The unhealthy patients who are detected in the misclassifier,

TN (True Negative) – The healthy patients who are detected in the classifier.

Based on these numbers we define the metrics as follows:

Accuracy: It means to find the ratio of the correct predictions to the total number of predictions and it is named as Exactness which is also called as Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{6}$$

F1-Score: The fair balance between the recall and precision is done statistically and stored in a value called as F1-Score

$$\mathrm{F1-Score} = 2 \times \frac{\mathrm{Precision} + \mathrm{Recall}}{\mathrm{Precision} \times \mathrm{Recall}} \qquad (7)$$

Precision: This is a ratio of truly detected positives to the total number of positives achieved in the classifier

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (8)$$

Recall: This defines the percentage of total positives detected correctly in the model statistically.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (9)$$

There are two scenarios that we consider for training our model.

1. **The patient-dependent scenario;** in this example, we train and test a model independently for each patient. The model is trained on the first N-10 data of a specific patient and evaluated on the final 10 measurements. The average RMSE over all patients is then used as an assessment statistic for a specific model. Because each patient does not have the same number of gathered measurements, the quantity of data required for training differs per patient.
2. **The patient-independent scenario**, in which we train a model on 54 patients before validating it on the remaining 6 cases. We next do cross-validation on the remaining 9 folds and utilize the average RMSE over these folds as our assessment measure for a specific model. We employ this method to get highly significant findings (since we can calculate the RMSE with 9 samples) despite being able to use a different validation set. This helps us to get the most out of the small quantity of data available. As more patients continue to join the randomized trial, they may be used as an extra test set in the near future.

Figure 3 *Old peak (Exercise-induced ST depression)*

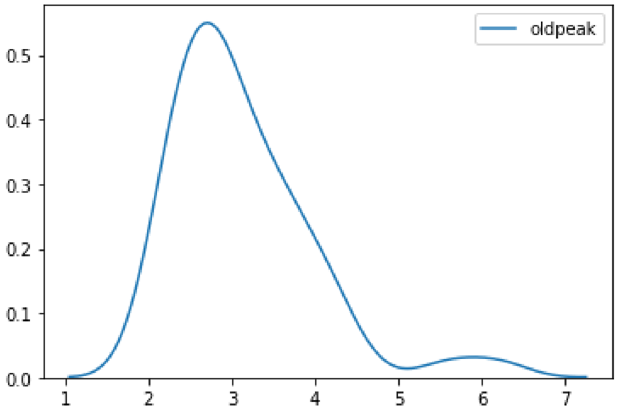
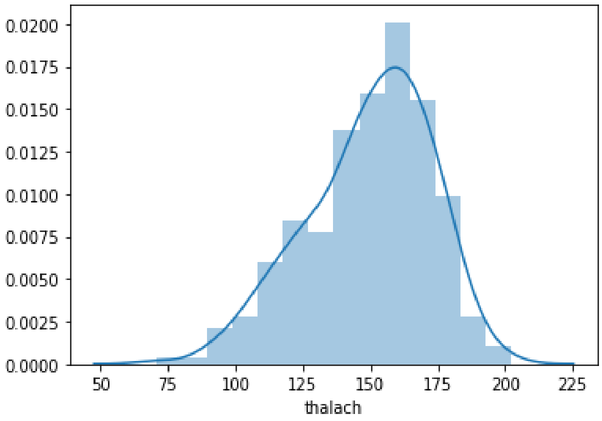


Figure 4 *thalach (Maximum heart rate)*

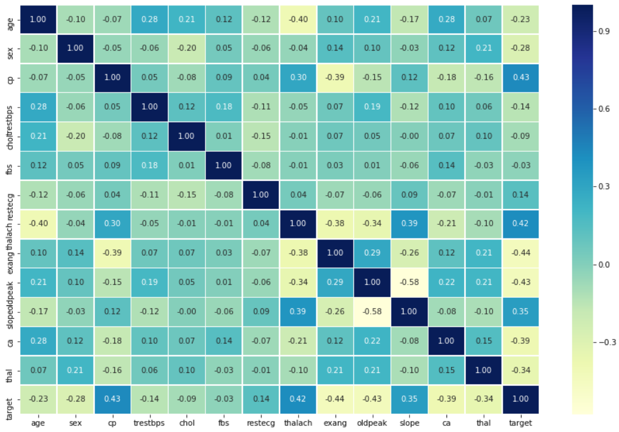


To identify the features of relevence heat map is created. The heat map shows that Chest Pain kind, Maximum heart rate , and peak of the slope has been most influence (more than 0.3) on predicting the CVD dataset (see Figure 5).

Table 1: Classification results three ML models and one DL model without feature selection

ML prototype	Exactness(accuracy)	Clarity (precision)	Recall	F1-score
KNN	0.65	0.68	0.71	0.75
LR	0.86	0.88	0.85	0.87
RF	0.75	0.76	0.77	0.78
Proposed work	0.91	0.92	0.91	0.92

Figure 5 *Heatmap of CVD dataset.*

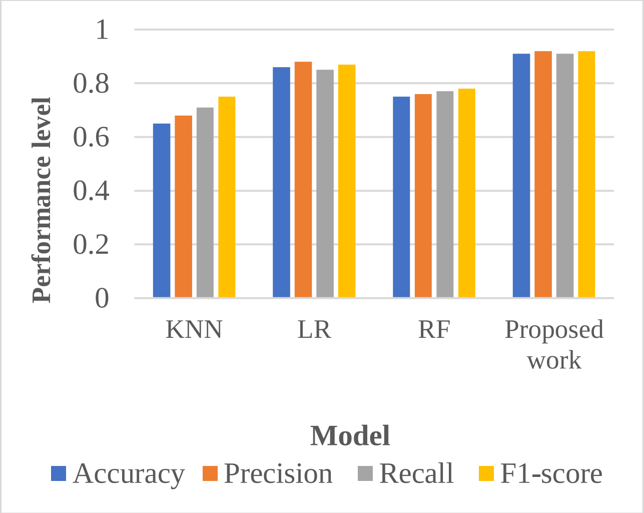


Our proposed work obtained a high exactness of 0.91, clarity of 0.92, recollect 0.91, the FI score is around 0.92 without using feature selection., yet Random Forest underperformed in all four criteria (see Figure 6). Our proposed approach differs from KNN, LR, and RF in that it integrates several layers in neural networks so that the mapping method called as non linear technique is used to extract input data with higher level features. The consequence here is it is not over fit and can be used straight forward.

Table 2: Classification results three ML models and one DL model with feature selection

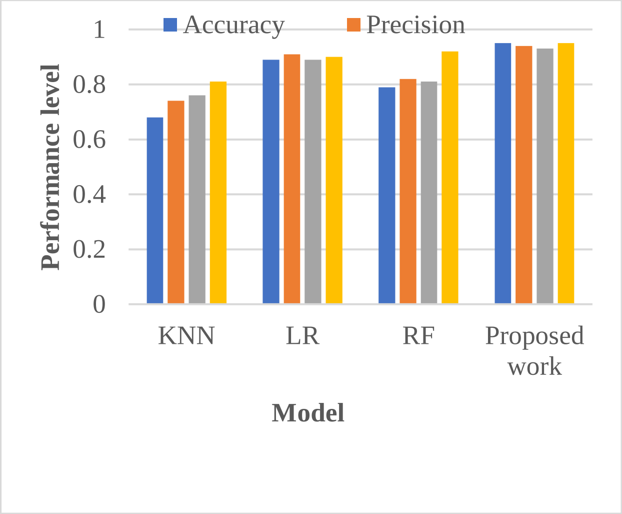
ML prototype	Exactness(accuracy)	Clarity (precision)	Recall	F1-score
KNN	0.68	0.74	0.76	0.81
LR	0.89	0.91	0.89	0.90
RF	0.79	0.82	0.81	0.92
Proposed work	0.95	0.94	0.93	0.95

Figure 6 *(BDLSTM) and KNN, LR, RF comparison by the proposed models when feature selection is not considered.*



Our suggested study produced high exactness of 0.95, clarity of 0.94, recollect 0.93, the FI score is around 0.95 after using feature selection. However, Random Forest underperformed in all four criteria (see Figure 7). Our proposed approach differs from KNN, LR, and RF in that it integrates several layers in neural networks so that the mapping method called as non linear technique is used to extract input data with higher level features. The consequence here is it is not over fit and can be used straight forward.

Figure 7 *Working comparison of metrics for proposed model (BDLSTM+Catboost) and KNN, LR, RF models with feature selection*



On top two features featured above in obtained datasets, The performance of the three ML classifiers are examined and recommended work (BDLSTM+Catboost). CP (score = 0.43), thalach (score = 0.42), and slope (score = 0.35) were the features chosen.

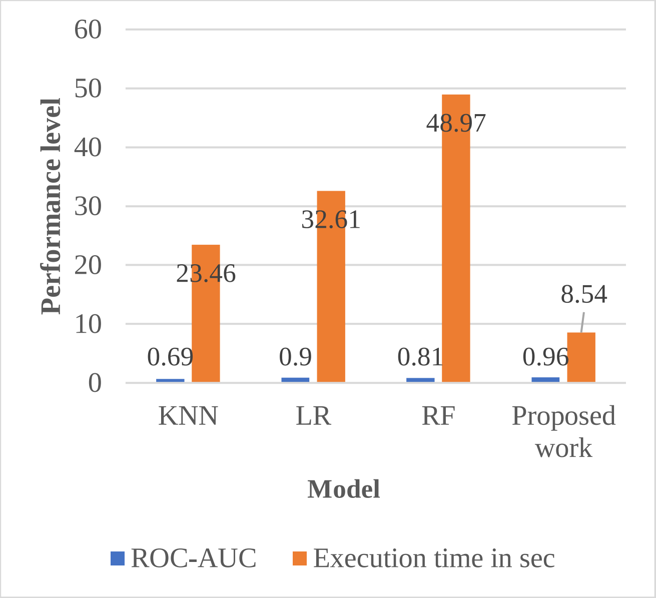
Table 3 displays the results of the four classifier models for the CVD dataset scenarios. When compared to the KNN, LR, and RF models, the results showed that BDLSTM+Catboost (proposed work) achieved the highest AUC (0.96). As we need some epochs to achieve results and have the quickest ET of 854 milliseconds as shown in the figure.

Table 3: The CVD datasets have top three attributes at they are classified into train and test set for performance analysis

Model	ROC-AUC	Execution time in sec
KNN	0.69	23.46
LR	0.90	32.61
RF	0.81	48.97
Proposed work	0.96	8.54

Figure 8

Performance analysis of classifiers methods on CVD dataset



5. Conclusion

The volume of healthcare data is always increasing at an alarming pace through various unreliable data sources. A streaming computing platform is required for improving patient outcomes and developing scalable real-time health status prediction systems. The suggested approach was utilized to create a CVD prediction model and to analyze treatment response using potential information confirmed instances. In terms of technology, our suggested framework aids in determining which algorithm is most suitable for usage in the pharmaceutical business in order to produce a device or software program that detects CVD-related issues at an early stage. The BDLSTM learning model outperformed the ML models in terms of predicting CVD incidence. Furthermore, Catboost was used to check that the known risk variables found to be relevant in prior clinical research were retrieved from the study findings. This might aid in the early diagnosis of people with diabetes and cardiovascular disease who could benefit from preventative therapy, reducing future healthcare costs.

References

[1] Abdalrada, A. S., Abawajy, J., Al-Quraishi, T., & Islam, S. M. S. (2022). Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study. Journal of Diabetes & Metabolic Disorders, 21(1), 251–261. https://doi-org.proxy1.library.eiu.edu/10.1007/s40200-021-00968-z.

[2] Ahishakiye, Emmanuel & Mwangi, Waweru & Muthoni, Petronilla & Nderu, Lawrence & Wario, Ruth. (2021). Comparative Performance of Machine Leaning Algorithms in Prediction of Cervical Cancer.

[3] Cheddad, Abbas. (2020). Machine Learning in Healthcare.

[4] Barbieri, S., Mehta, S., Wu, B., Bharat, C., Poppe, K., Jorm, L., & Jackson, R. (2022). Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. International Journal of Epidemiology, 51(3), 931–944. https://doi-org.proxy1.library.eiu.edu/10.1093/ije/dyab258.3

[5] Dorogush, Anna & Ershov, Vasily & Gulin, Andrey. (2018). CatBoost: gradient boosting with categorical features support. arXiv:1810.11363

[6] Ed-daoudy, A., Maalmi, K. A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment. J Big Data 6, 104 (2019). https://doi.org/10.1186/s40537-019-0271-7.

[7] Graves, Alex & Schmidhuber, Jürgen. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks: the official journal of the International Neural Network Society. 18. 602-10. 10.1016/j.neunet.2005.06.042.

[8] Habehh, Hafsa & Gohel, Suril. (2021). Machine Learning In Healthcare. Current Genomics. 22. 10.2174/1389202922666210705124359.

[9] Keto, Jaana & Ventola, Hanna & Jokelainen, Jari & Linden, Kari & Keinänen-Kiukaanniemi, Sirkka & Timonen, Markku & Ylisaukko-oja, Tero & Auvinen, Juha. (2016). Cardiovascular disease risk factors in relation to smoking behaviour and history: a population-based cohort study. Open Heart. 3. e000358. 10.1136/openhrt-2015-000358.

[10] Olokoba, Abdulfatai & Olusegun, Obateru & Lateefat, Olokoba. (2012). Type 2 Diabetes Mellitus: A Review of Current Trends. Oman medical journal. 27. 269-73. 10.5001/omj.2012.68.

[11] Puaschunder, Julia. (2020). The Potential for Artificial Intelligence in Healthcare. SSRN Electronic Journal. 10.2139/ssrn.3525037.

[12] Li, Bin & Yu, Qingzhao & Peng, Lu. (2019). Ensemble of fast learning stochastic gradient boosting. Communications in Statistics - Simulation and Computation. 1-13. 10.1080/03610918.2019.1645170.

[13] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

[14] Rustam, F., Ishaq, A., Munir, K., Almutairi, M., Aslam, N., & Ashraf, I. (2022). Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction. Diagnostics (2075-4418), 12(6), 1474. https://doi-org.proxy1.library.eiu.edu/10.3390/diagnostics12061474.

[15] Sidey-Gibbons, A. M. Jenni, and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” BMC Medical Research Methodology, vol. 19, 2019.

[16] Sung, J. M., Cho, I.-J., Sung, D., Kim, S., Kim, H. C., Chae, M.-H., Kavousi, M., Rueda-Ochoa, O. L., Ikram, M. A., Franco, O. H., & Chang, H.-J. (2019). Development and verification of prediction models for preventing cardiovascular diseases. PLoS ONE, 14(9), 1–12. https://doi-org.proxy1.library.eiu.edu/10.1371/journal.pone.0222809.

[17] Tekkeşin, Ahmet. (2019). Artificial Intelligence in Healthcare: Past, Present and Future. The Anatolian Journal of Cardiology. 22. 10.14744/AnatolJCardiol.2019.28661.

[18] The Top 10 Causes of Death. Available online: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed on 31 May 2021).

[19] Thireou, Trias & Reczko, Martin. (2007). Bidirectional Long Short-Term Memory Networks for Predicting the Subcellular Localization of Eukaryotic Proteins. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM. 4. 441-6. 10.1109/tcbb.2007.1015.

[20] Xie, J., Wu, R., Wang, H., Chen, H., Xu, X., Kong, Y., & Zhang, W. (2021). Prediction of cardiovascular diseases using weight learning based on density information. Neurocomputing, 452, 566–575. https://doi-org.proxy1.library.eiu.edu/10.1016/j.neucom.2020.10.114.