Heart Disease Prediction Integrating UMAP and XGBoost

Ayushi, Shilpa Sethi, Jyoti

Abstract*: *The healthcare industry is flooded with the plethora of data about the patients which is supplemented each day in the form of medical records. Researchers have been putting in various efforts to bring this data into usage for the prediction of various diseases. Prediction of heart diseases is one such area. Data mining algorithms have been at the centre of improving the prediction of accuracy of heart diseases. But it has been found that these algorithms are not using adequate set of attributes for prediction that sometimes may lead to wrong predictions. The aim of this paper is to deploy the right set of algorithms to accurately predict the heart diseases and help both the patient and the doctor. The paper thrives to put UMAP and XGBoost techniques in this regard and exploit the advantages of both techniques. UMAP helps in dimensionality reduction without loss of useful data while XGBoost uses parallelization for tree construction reducing the time required to get the results. The experiment is carried on real data taken from Fortis Escorts, Faridabad, India. The results are compared with existing techniques such as Naïve Bayes, Decision Tree model, Logistic Regression model and Support Vector Machine (SVM) model based on various parameters such as accuracy, recall and precision. Remarkable accuracy of 94.59%, recall of 87.87, precision of 100 has been achieved.

Keywords*: *Classification algorithms, Ensemble Techniques, PCA, UMAP, XGBoost.

I. INTRODUCTION

According to World Health Organization 31% of global deaths are caused by Cardio Vascular Diseases (CVD) [1], 80% of which are due to heart attack and stroke [2]. It is expected that by 2030 approximately 23.6 million individuals will decease due to Heart malady [3]. As per Global Burden of Disease Report, released on September 15 2017, 1.6 million people in India died of heart disease in 2016 [4]. University of Rochester's Medical Centre claims that major causes of heart disease are lack of physical activity, obesity, tobacco consumption, smoking and alcohol [5]. Many CVDs can be prevented if behavioural risk factors such as physical inactivity, tobacco use, obesity, unhealthy diet and harmful use of alcohol are under control [1]. Others having some risk factor such as diabetes, hypertension or hyperlipidaemia may be prevented by identifying the disease at early stage and taking proper precautions for recovery [6]. But it has been observed that the clinical decisions are often influenced by medical errors, unwanted medical costs and personal bias at doctor’s end [7], [8]. Therefore, training a machine for predicting existence of heart disease using machine learning algorithms becomes necessary in addition to using the expertise of doctors. Various efforts have been put by researchers in this domain like authors [9], [10] used association rule mining and classification rule mining to study the patterns in data, information mining is used in [11] while [12], [13] proposed decision support using Naïve Bayes to build an accurate model to predict heart disease. Simple classification algorithms have also been used by researchers [14] to build their model. Variable Centered Intelligence Rule System in [15] and Multiple Linear Regression & Hybrid Genetic Algorithm in [16] were used.

All the techniques involved inconsistencies at some point of time either in accuracy levels or methodology. Large costs were involved in some, while some researchers did not have adequate data to support their findings. To overcome these problems, amalgam of UMAP with XGBoost is proposed in this paper. The main objectives of this work are to develop an intelligent heart disease diagnosis mechanism and study the relevancy of attributes that may cause heart disease. In UMAP-XGB, firstly the data collected from different sources is pre-processed using various tools in python. Once the data is processed UMAP is used for dimensionality reduction and based on the machine feature set generated by UMAP a model using XGBoost is built. Tuning of XGBoost model is done after which it is tested and evaluated. Accuracy of 94.59% has been achieved by the proposed model.

This paper is organized as follows: Section II covers the related work and Section III describes the proposed technique in detail. Detailed algorithms and flowchart have also been included. Section IV gives the experimental results and compares the proposed technique with some other most prevalent techniques of similar area. Conclusion and future scope are summarized in Section V.

II. LITERATURE REVIEW

A. H. Chen [17] developed a heart disease prediction system in 2011 by selecting 13 most effective parameters from the dataset. An artificial neural network was developed for classification with predictive accuracy of 80%. Authors [18] developed a heart disease decision support system.

Manuscript received on April 30, 2020.
Revised Manuscript received on May 06, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author

Ayushi*, Department of Computer Science, J. C. Bose University of Science and Technology, YMCA, Faridabad, India. Email: ayushibansal702@gmail.com

Shilpa Sethi, Department of Computer Applications, J. C. Bose University of Science and Technology, YMCA, Faridabad, India. Email: Munjal.shilpa@gmail.com

Jyoti, Department of Computer Science, J. C. Bose University of Science and Technology, YMCA, Faridabad, India. Email: justjyoti.verma@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

A Naïve Bayesian classification algorithm was developed in order to extract hidden patterns from heart disease patient dataset. A web-based application was developed which enabled users to answer a set of predefined questions [19].

H.D. Masethe and M. A. Masethe [14] claimed to achieve a significant accuracy ~99% but it has been argued that many of the relevant attributes were missing and experiment was being conducted using limited dataset.

Further, an integrated approach using K-nearest neighbour and Ant Colony Optimization was proposed in [20]. The technique only achieved an accuracy of 70.26% with error rate of 0.526. Authors in [21] used ensemble techniques bagging, boosting and staking. Six different classification algorithms namely Bayes Net, Multilayer perceptron, Naïve Bayes, Decision tree classifiers and Sequential Minimal Optimization for the support vector machine were used with the ensemble techniques. The results are analysed using WEKA tool. The study showed that Naïve Bayes resulted in most accuracy with every technique and achieved 85.3% accuracy for Heart Disease dataset and 92% for Heart Valve Disease dataset.

Different rules such as Original Rules, Rules without duplicates, Pruned Rules, Sorted Rules, Classified Rules and Polish were generated using decision tree in [22] for heart disease prediction. Knowledge Extraction based on Evolutionary Learning was used to generate rules. Further, a technique with called Hybrid Classifier with Weighted Voting was proposed in [8]. It used nine classifiers namely SVM, Neural Network, Decision tree, generalized linear model, Lasso, Bayesian regularized Neural network, Classification and Regression Tree, Multivariate Adaptive Regression Spline to generate an ensemble and achieved an accuracy of 82.54%. Recently a technique based on Spark MLib and classification model was proposed in [23] to evaluate the performance of Random forest for predicting heart disease. Though the time taken by random forest to evaluate number of records was less compared to traditional data mining tool due to in- memory computations, the dataset used was taken from UCI repository diluting the purpose of using real time streaming tool.

A critical look at the literature survey indicates that ensemble of algorithms enhances the efficiency of any system over usage of single algorithm. The results are much less dependent on the peculiarities of single algorithm and it makes the model more expressive and less biased [24], [25]. Hybrid techniques becomes even more useful when data is collected from different sources in order to make informed decision [21].

Heart disease prediction has always been a hot topic among researchers but the available literature has not addressed some common challenges in traditional heart disease diagnosis systems such as large medical costs, unwanted bias at doctor's end and delay in predictions. Most of the research works used the data from UCI repository only and features are not studied properly. Some of the features in the dataset used are not important, hence unnecessarily add to the medical costs in addition to reducing the accuracy of model. Not only adequate features need to be considered, but the features under consideration should be optimized. All these issues should be raised and solved for an efficient heart disease prediction system.

III. METHODOLOGY

Owing to the various problems identified there arise a need to form a medical database that contains real time data of patients. Data is collected from hospitals [26] in addition to UCI repository [27]. The aim is to mirror the current pattern of features that lead to heart disease. After populating the database irrelevant features are identified using Information value [28] and dropped. This results in reduction in medical costs as greater number of features requires a greater number of tests to be carried. Medical costs can be easily cut if these features are identified and discarded. Driven by the challenges a hybrid technique using UMAP [29] and XGBoost is proposed in this paper. On selecting important features UMAP is used to reduce the dimensionality of the data further. This is done in order to reduce the number of features in the feature set on which the model is to be trained without losing useful data. This helps in reducing the time and storage space required for the model so that large amount of data can be handled efficiently. XGBoost is used to build a model using the machine feature set obtained from UMAP. The proposed framework for Heart Disease Prediction Integrating UMAP and XGBoost is shown in Fig. 1

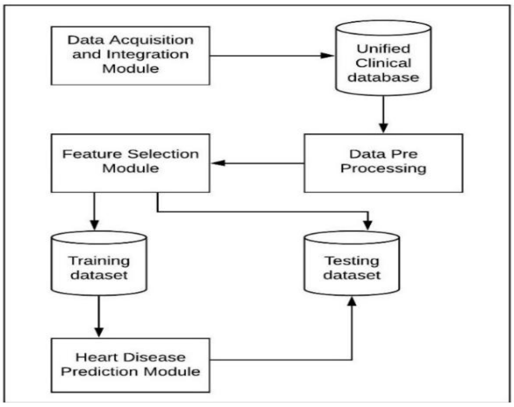
Data Acquisition and Integration Module
Unified Clinical database
Feature Selection Module
Data Pre Processing
Training dataset
Testing dataset
Heart Disease Prediction Module

Fig. 1. HDPS- Heart Disease Prediction System

The detail description of each module is given in following subsections.

A. Data Acquisition and Integration module

Data acquisition aims to collect data relevant to heart disease. Collection of patient data is a crucial step for any work done in medical domain. Datasets pertaining to heart patients are freely available on UCI repository [27]. Since the repository is ~32 years old there is a need to capture the change in features mounted in the repository with passing years. A validation check needs to be performed in order to know if the features are appropriate so there is a strong need to augment this with real time data. Thus, emphasis was laid on collecting data from the hospitals [26] by consulting the domain expert (Senior Cardiologist). Data integration focuses on integrating data obtained from various sources and store it in a uniform format in Unified clinical database. The algorithm for data integration is given below.

```
DI( )
Input: Multiple datasets
Output: Unified dataset
1. for each dataset
2. form a dataframe $D_i$
3. do
4. Append Di at the end of compiled dataframe $D_c$
```

```
for each attribute in $D_c$
Check if $D_i$ has same attribute
if missing attribute
continue
while ($D_i$)
i=i+1
```

B. Unified Clinical Database

Unified clinical database is the integrated database from all sources. All the patient records collected are stored in this database. The attributes used and their possible values are described in Table I.

Table- I: Dataset description

Risk Factor	Possible values	Description
Age	Integer	Displays the age of the individual
Sex	0 = female, 1= male	Displays the gender of the individual
Chest-pain type (cp)	1 = typical angina, 2 = atypical angina, 3 = non angina pain, 4 = asymptotic	Displays the type of chest pain experienced
Resting blood pressure	Integer	Displays resting blood pressure value in mmHg
Serum Cholesterol (chol)	Integer	Displays the serum cholesterol in mg/dl
Fasting Blood Sugar (fbs)	fbs > 120 = 1 fbs < 120 = 0	Compares fasting blood sugar value of an individual with 120 mg/dl
Resting ECG	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hyperthrophy	Displays resting electrocardiographic results
Max heart rate achieved	Integer	Displays maximum heart rate achieved by an individual
Exercise induced angina	0 = no, 1 = yes	
ST depression (oldpeak)	Integer or float	Induced by exercise relative to rest
Slope	1 = up slope, 2 = flat, 3 = down slope	Peak exercise ST segment
Number of major vessels coloured by fluoroscopy	Integer or float	
Thalassemia (thal)	3 = normal, 6 = fixed defect, 7 = reversible defect	
Target	0 = absence, 1,2,3,4 = presence	Displays whether the individual is suffering from heart disease or not

C. Data Pre-Processing

Data pre-processing refers to preparing data to make it suitable for data mining. The data collected is prone to outliers and missing values. Since some techniques are sensitive towards missing values and may not work till these are treated, there is a need to impute the missing values. Missing data further affects the accuracy of model.

For treating missing values, data is analysed to see if the features are continuous or categorical. If the features are continuous, outliers are taken into consideration. Here, outliers refer to data points that are distant from other similar data points. Outliers are visualized using Boxplot [30].

If the continuous feature has outliers too far from other values, mode of feature is used to fill the missing fields else mean of the features is calculated and used in place of missing values. In case of categorical features, the missing fields are filled forward or backward. The algorithm for Data pre-processing module is given below.

```
DP( )
Input: Unified dataset
Output: Processed data
while ($D_c$) {
analyse data
for each attribute $A_i$
if $A_i$ categorical
fillna(pad)
else
if outlier
fillna(mode)
else
fillna(mean)
```

Outliers in the underlying database are depicted in Fig. 2.

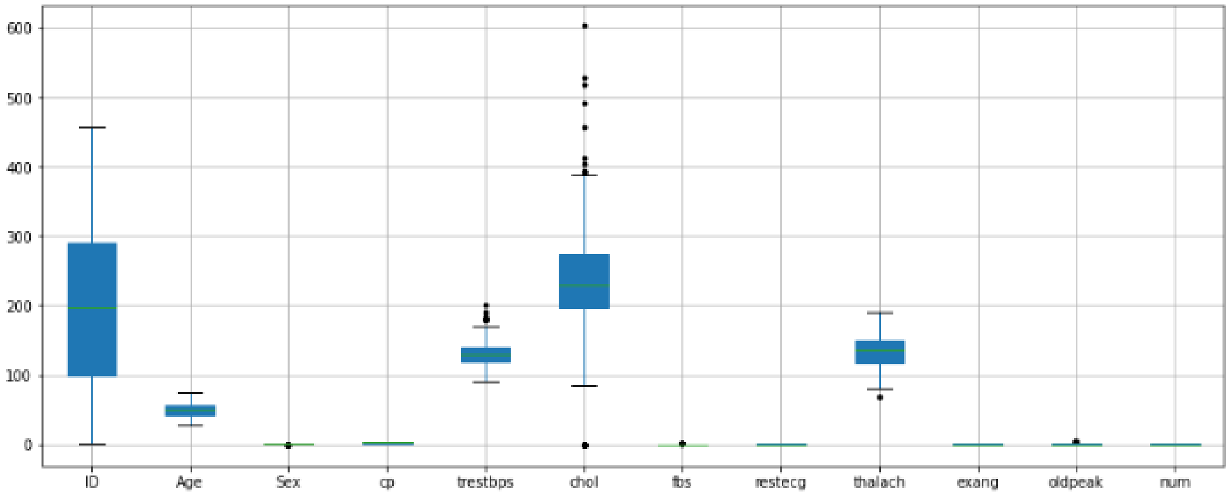

Fig. 2. Outliers in Unified Clinical database

D. Feature Selection Module

The collected data is classified into set of various attributes. Choice of attributes depends on its contribution to heart disease. It has been observed that various methods such as correlation, logistic regression, information value can be applied on our dataset to find the relationship between dependent and independent attributes. The relationship helps in determining which attribute is important for the target value and which do not contribute much. Information value (*IV*) is used in this work which is computed using Weight of Evidence (WOE). WOE tells the predictive power of an independent variable with respect to dependent variable and describes the relationship between these variables while *IV* measures the strength of relationship between independent and dependent variables. It is related to sum of WOE over all groups.

WOE measures strength of bins and separate events from non-events by computing the ratio given in equation (1).

$$R = \frac{Distribution\ of\ Events}{Distribution\ of\ NonEvents} \qquad (1)$$

WOE and IV are calculated using formulas given in equations (2) and (3).

$$WOE = ln\left(\frac{Event\ \%}{Non\ Event\ \%}\right) \qquad (2)$$

where

- Event is presence of heart disease
- Non-Event is absence of heart disease

$$IV = \sum(Event\% - Non\ Event\%) * WOE \ (3)$$

Attributes are selected using the Information value according to Table II.

Table II: Information value table

IV	Predictive Power
<0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
>0.5	Suspicious or too good

The algorithm for feature selection module is given.

```
FSM( )
Input: Processed data
Output: Reduced dataset
1. do
2. for each independent variable
3. construct bins
4. while (bins)
5. calculate WOE for each bin
6. calculate IV
7. sort values based on IV
8.remove the independent variable with lower IV
```

E. Heart Disease Prediction module

The major contribution of this paper is achieved by Heart Disease Prediction module. Aim of this module is to predict the presence or absence of heart disease. Various techniques may be chosen to build a model for prediction such as Fuzzy rule base, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Naïve Bayes and XGBoost. Different experiments are carried based on which different observations are recorded which have been summarized in Table III.

Table- III: Observations based on recent techniques used

Recent Techniques	Observation
Fuzzy Rule Base	The results of fuzzy rule base are based on assumption as these systems do not have the capability of learning as machine learning algorithms or of pattern recognition. It was observed that rule base cannot be formed without manually forming the rules, hence did not resolve the challenge of introduction of personal bias.
RNN	As RNN is a sequential model, accuracy obtained by this model is low. The long-term data has to travel through all the units before reaching the processing unit, increasing the chances of corrupting the data by being multiplied several times by a small number, i.e. vanishing gradient is introduced.
LSTM	LSTM is considered to remember any new attribute in the input data leading to heart disease. It handled the vanishing gradient to some extent by using switch gates, but did not remove it completely. It is observed that there exists a sequential path from older units to the processing units.
Naïve Bayes	Naïve Bayes assumes features to be independent of each other but the patient dataset does not fulfil this criterion and hence Naïve Bayes could not be used for building the required model.
XGBoost	It was observed that the model is well suited for our data. XGBoost has in-built regularization and hence overfitting is prevented in the model. Also, it uses the concept of parallelization improving speed.

All the above-mentioned techniques are implemented and trained using the underlying database. It has been observed that XGBoost offers better speed when compared to other methods and is well suited for our data. Hence, we consider XGBoost for predicting the heart disease. The objective function used in our proposed model is given in equation (4)

$$Obj(\theta) = L(\theta) + \Omega(\theta) \qquad (4)$$

where

- θ is the model parameter
- $L(\theta)$ is training loss
- $\Omega(\theta)$ is regularization term

Ω represents complexity of the model and L represents matching degree between model and training set.

The next challenge after finding a suitable technique for building the model is to reduce the dimensionality of data. Here the major focus is to reduce the dimensionality of data without loss of useful data. The multi-collinearity of the features is removed. Reducing the dimensions also lead to a considerable reduction in execution time and storage space required by the model. The two most prevalent techniques namely Principal Component Analysis (PCA) and UMAP are considered for this purpose.

Though both techniques try to preserve the most of useful information, it is found that PCA is a linear transformation technique and needs a careful selection of number of Principal components. If the principal components are not chosen carefully, it misses some useful information compared to original data. In contrast, UMAP captures the non-linearity among features. It is used over PCA as it preserves the global structure of data better. Exponential probability distribution is used by UMAP for dimensionality reduction. Probability function of UMAP is given by equation (5)

$$P_{i|j} = e^{\frac{-d(x_i,x_j)-\rho_i}{\sigma_i}} \qquad (5)$$

where

- ρ is distance between i^{th} data point and its nearest neighbour.

Algorithm for Heart Disease Prediction module.

HDPM()
Input: dataset with d dimensions
Output: trained model

```
1. construct a reducer R for UMAP
2. fit an object E with R on $D_p$
3. set optimal value for n_neighbors n
4. set min_dist m
5. transform E using m,n
6. split $D_p$ into independent, dependent variables X,y
7. fit X using R
8. split X, y in train and test data
9. build XGBoost model M
10. train M on train dataset
11. evaluate M on test dataset
12. for M compute accuracy, recall and precision
```

IV. RESULTS AND DISCUSSION

A. Experimental Setup

Python 3.0 is used for building the model by using Jupyter notebook, which is an interactive notebook and has an integrated development environment for machine learning applications. This environment supports Python scripts and is used to perform data analysis and model building.

B. Performance Metrics

Classification report and Area Under the Curve-Receiver Operating Characteristics (AU-ROC) Curve are used to evaluate the performance of proposed model.

***Classification Report**:* The model is evaluated using classification report which is used to measure the quality of predictions of a classification algorithm using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) variables. It shows precision, recall and f1-score on per class basis.

Precision: Here, precision is the percentage of correct predictions, i.e. accuracy of positive predictions. It is defined as the ratio of true positives to the sum of true and false positives and is given in equation (6)

$$Precision = TP/(TP + FP) \qquad (6)$$

Recall: Here, recall is the ability of a classifier to find all positive instances, i.e. it is the fraction of positives correctly identified.

It is defined as the ratio of true positives to the sum of true positives and false negatives. Recall is calculated using equation (7)

$$Recall = TP/(TP + FN) \qquad (7)$$

F1 score: F1 score in the context is defined as the percent of correct positive predictions. It is a weighted harmonic mean of precision and recall and is given by equation (8)

$$F1\ Score = \frac{2*(Recall*Precision)}{(Recall+Precision)} \qquad (8)$$

AU-ROC: It is used to measure the performance of model for classification problem at various thresholds settings. ROC is a probability curve while AUC represents the degree or measure of separability. It tells how much a model is capable of differentiating between classes. Higher the AUC, better

Table IV gives the comparative analysis between these techniques and proposed technique.

the model is at differentiating between patients with disease and no disease. The curve is plotted with True Positive Rate (TPR) against False Positive Rate (FPR). TPR and FPR are given by equations (9) and (10)

$$TPR = TP/(TP + FN) \qquad (9)$$

$$FPR = FP/(TN + FP) \qquad (10)$$

The model is compared to some other powerful techniques namely Logistic Regression, Naïve Bayes, SVM, Random forest and XGBoost. Table gives the evaluation of UMAP-XGB with above mentioned techniques based on Accuracy, Precision, Recall and Area under ROC curve.

Table- IV: Comparative evaluation of various techniques

Technique	Accuracy	Precision	Recall	Area under ROC Curve
Logistic Regression	85.5	79.16	79.16	84.02
Naïve Bayes	84.05	87.5	72.41	84.86
Support Vector Machine	86.95	91.66	75.86	88.05
Random forest	85.5	79.16	86.36	84.02
XGBoost	86.95	83.33	80	86.11
UMAP-XGB	**94.59**	**100**	**87.87**	**95.56**

ROC curves for various techniques are given in Fig. 3.

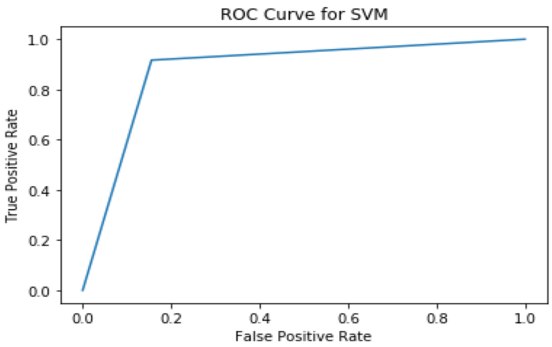

Fig. 3(a). SVM ROC Curve

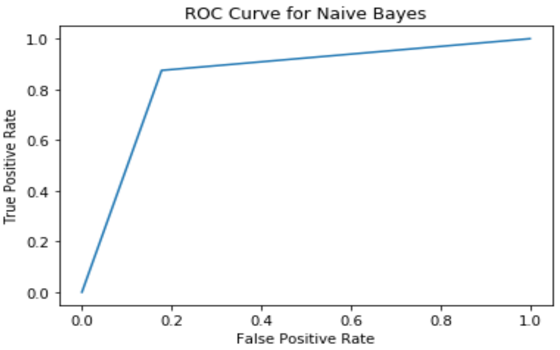

Fig. 3(b) Naïve Bayes ROC Curve

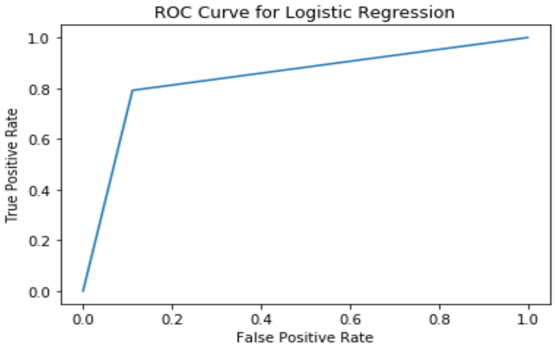

Fig. 3(c). Logistic Regression ROC Curve

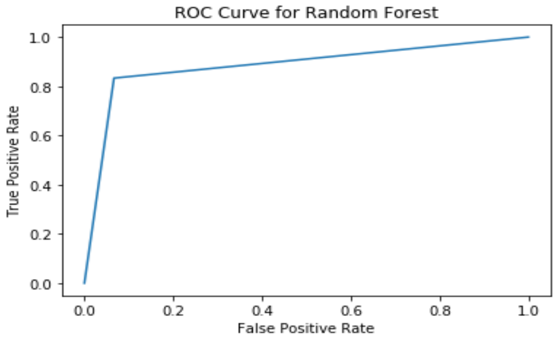

Fig. 3(d). Random Forest ROC Curve

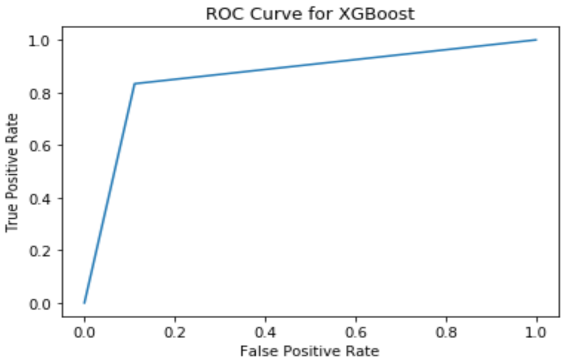

Fig. 3(e). XGBoost ROC Curve

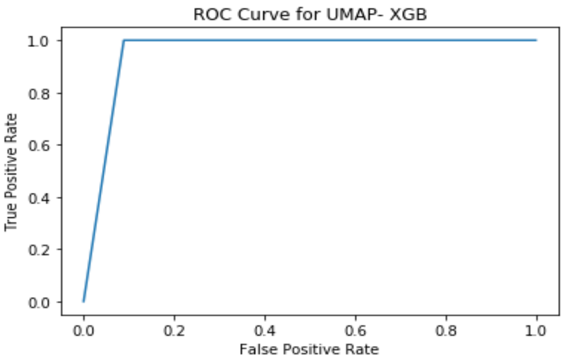

Fig. 3(f). UMAP-XGB ROC Curve

Accuracy, precision and recall of all above mentioned techniques are shown in Fig. 4, 5 and 6.

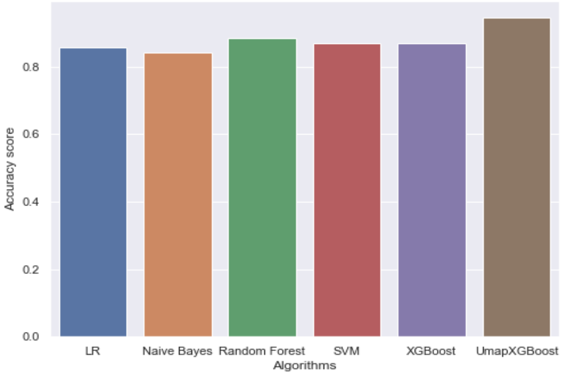

Fig. 4. Comparison based on accuracy

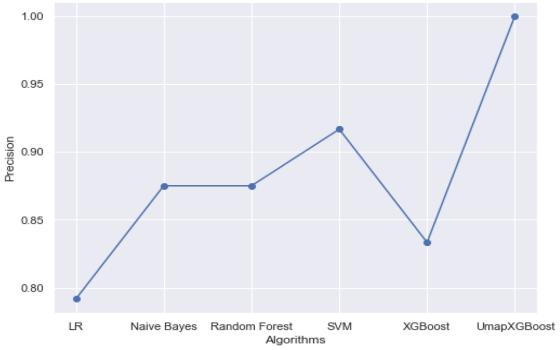

Fig. 5. Comparison based on Precision

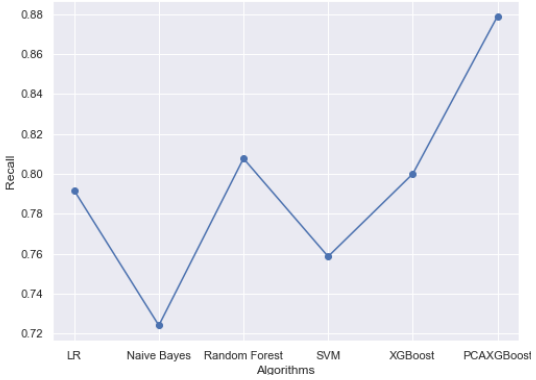

Fig. 6. Comparison based on Recall

PCA and UMAP have been applied to same model built using XGBoost and the results are shown in Fig. 7

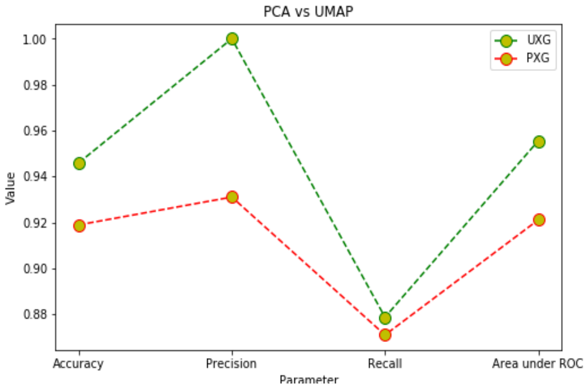

Fig. 7. PCA vs UMAP

It can be observed from the figure that UMAP fits the data better than PCA when its machine feature set is applied to XGBoost model for predicting heart disease. The model trained using UMAP has higher accuracy, precision, recall and area under ROC curve.

V. CONCLUSION AND FUTURE SCOPE

With increasing number of heart-patients the size of database is increasing tremendously but all the data is not maintained properly to draw analysis from it. So, it has been difficult for researchers to obtain value from data. All the researches have been conducted on a single database that use set 13 attributes from UCI repository leading to increased costs due to greater number of attributes. Many researches have been conducted to build an efficient system for heart disease prediction but there exist many key challenges that have been left untouched and that may improve the quality of heart disease prediction to a great extent. Towards this goal, an extensive literature survey has been done and challenges have been identified. Driven by the challenges a framework has been proposed that focuses on working towards the challenges identified.

The major objective of the work was to build an efficient heart disease prediction system that identifies the risk accurately. To achieve the objective, a model using hybrid of UMAP and XGBoost has been proposed in this dissertation. The major contributions of this work are listed below.

A. Hybrid of UMAP-XGB

Ensemble of algorithms enhances the efficiency of any system over usage of single algorithm as the results are much less dependent on the peculiarities of single algorithm and it makes the model more expressive and less biased. It becomes even more useful as the data is collected from multiple sources.

B. Improved Accuracy

Accuracy of the model plays a vital role in early heart disease prediction models. Another major contribution of this work is its major improvement in accuracy. Accuracy has been increased to 94.59% compared to recent researches (accuracy ~82% to ~87%) which is a crucial aspect for disease prediction systems.

C. Improvement in Efficiency

The proposed framework results in reducing the execution time of the model due to parallelization used by XGBoost, enhancing the efficiency of model. Also, storage space required by the model is minimized and that helps in handling large datasets proficiently.

D. Reduced Medical Costs

Three features out of 13 have been identified that do not contribute much in prediction. This is a major contribution in reducing the medical costs involved in prediction of heart disease as larger number of features indicate larger number of medical tests.

E. Elimination of Personal Bias

The model is trained using machine learning algorithms so there are no chances of inclusion of factors such as intuition and personal bias. The knowledge is derived out of medical databases.

F. Improved Results

Accuracy achieved = 94.59%
Improvement in Accuracy = ~8%
Precision achieved = 100
Improvement in Precision = ~9%
Recall achieved = 87.87
Improvement in Recall = ~1%
Area under ROC Curve = 95.56
Improvement in Area under ROC curve = ~7%

G. Results per class basis

For absence of heart disease (class 0)

Precision achieved = 100
Recall achieved = 91
F1 score achieved = 95

For presence of heart disease (class 1)

Precision achieved = 88
Recall achieved = 100
F1 score achieved = 94

Result analysis proves that the proposed framework outperforms other works on each aspect and is a suitable model was heart disease prediction. The proposed model is reducing cost as well as performing better prediction.

The proposed technique performs good on every aspect but needs a manual input to be fed to determine the risk of heart disease. There must be a provision to automatically detect the values of features required to determine the possibility of disease. An alert system should be embedded with the model that can alarm the user in real time. Also, the model is trained on structured data. Image processing techniques can be used so that values can be read from image scans.

The work can be extended in future by

- embedding nano-technology and use of sensors in order to fetch data instantly that can alert the patient in real time.
- Enabling the model to receive inputs from an image and getting the required values through image processing

REFERENCES

1. World Health Organization, *Cardiovascular Diseases (CVDs)* viewed 14 April 2020, https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Hazra, S. Mandal, A. Gupta, and A. Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review," *Advances in Computational Sciences and Technology*, 2017, vol 10, pp. 2137-2159
3. Yusheng Fu, and Jinhong Guo, "Blood Cholesterol Monitoring with Smartphone as Miniaturized Electrochemical Analyzer for Cardiovascular Disease Prevention", *IEEE Transactions on Biomedical Circuits and Systems*, 2018, vol. 12, no. 4, pp. 784-790.
4. H. Jayasree, D. S. S. K. R. T. Naren, K. Sai Sankeerth, T. Kumidini, “Heart Disease Prediction System (HDPS)”, *Journal of Applied Science and Computation (JASC)*, 2019, vol 6, no. 6, pp 2168-2175.
5. Yeshvendra K. Singh, Nikhil Sinha., and Sanjay K. Singh., “Heart Disease Prediction System Using Random Forest”, *First International Conference*, ICACDS 2016, pp 613-623.
6. Debabrata Swain, Santosh Kumar Pani, Debabala Swain. “An Efficient System for the Prediction of Coronary Artery Disease using Dense Neural Network with Hyper Parameter Tuning”, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2019, vol. 8, no. 6S.
7. Ranjana Joshi, and S. Sethi, “Prediction of Heart Disease Using Data Mining Technique”, *International Journal of Computer Sciences and Engineering*, 2018, vol. 6, Issue. 12, pp. 305-309.
8. M. Saini, N. Baliyan and V. Bassi, "Prediction of heart disease severity with hybrid data mining," *2nd International Conference on Telecommunication and Networks (TEL-NET), Noida, 2017*, pp. 1-6.
9. Sunita Soni, and O. P. Vyas, “Using Associative Classifiers for Predictive Analysis in Health Care Data Mining”, *International Journal of Computer Applications*, July 2010, vol. 4, no. 5.
10. Anushya & A. Pethalakshmi, “A Comparative Study of Fuzzy Classifiers with Genetic On Heart Data”, *International Conference on Advancement in Engineering Studies & Technology*, July 2012.
11. P. Dixit, S. S., A. K. Sharma and A. Dixit, “Design of an Automatic Ontology Construction Mechanism Using Semantic Analysis of the Documents”, *Fourth International Conference on Computational Intelligence and Communication Networks*, 2012, pp. 611-616.
12. Mariammal. D, Jayanthi. S, Dr. P. S. K. Patra, “Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques”, *International Journal of Advanced Research in Computer Science & Technology IJARCST*, 2014, vol. 2, no. Special 1.
13. Mrs. G. Subbalakshmi and Mr. K. Ramesh, “Decision Support in Heart Disease Prediction System using Naive Bayes*”, International Journal of Computer Science and Engineering (IJCSE)*, 2011, vol. 2, no. 2
14. H. D. Masethe, and M. A. Masethe, "Prediction of Heart Disease using Classification Algorithms", *Proceedings of the World Congress on Engineering and Computer Science*,2014, vol 2, San Francisco, USA.
15. F. Zennifa, Fitrilina, H. Kamil, and K. Iramina, "Prototype early warning system for heart disease detection using Android application", *36th Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2014, pp. 3468-3471.
16. K. Phiwhorn, and S. Arch-int, "LDL-Cholesterol Levels Measurement using Hybrid Genetic Algorithm and Multiple Linear Regression", *International Conference on Information Science and Applications,2013*, pp. 1-4.
17. A H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, “HDPS: Heart disease prediction system”, *Computing in Cardiology*, 2011, pp. 557-560.
18. S. A. Pattekari and A. Parveen, “Prediction System for Heart Disease Using Naive Bayes*”, International Journal of Advanced Computational and Mathematical Sciences.*, 2012, vol. 3, no. 3, pp. 290-294.
19. S. Sethi and Ashutosh Dixit, "Design of personalised search system based on user interest and query structuring," *2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015, pp. 1346-1351.

20. Rajathi, G. Radhamani, “Prediction and analysis of Rheumatic heart disease using KNN classification with ACO,” *International Conference on Data Mining and Advanced Computing (SAPIENCE), IEEE*, 2016, pp. 6873
21. R. El Bialy, M.A. Salama, O. Karam, “An ensemble model for Heart disease data sets: a generalized model,” *Proceedings of the 10th International Conference on Informatics and Systems*, 2016, pp. 191196.
22. Purushottam, Kanak Saxena, and Richa Sharma, "Efficient Heart Disease Prediction System", 2016, vol 85, pp. 962-969
23. A. Ed-daoudy, and K. Maalmi, “Real-time machine learning for early detection of heart disease using big data approach”, *International Conference on Wireless Technologies, Embedded and Intelligent Systems*, 2019, pp. 1-5.
24. G. Sujatha and K. U. Rani, “An Experimental Study on Ensemble of Decision Tree Classifiers*”, International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 2013, vol. 2, no. 8, pp. 300-306.
25. R. D. Kulkarni, “Using Ensemble Methods for Improving Classification of the KDD CUP ’99 Data Set”, *IOSR Journal of Computer Engineering*, 2014, vol. 16, no. 5, pp. 57-61.
26. Fortis Escorts, 2019, *Interventional Cardiology*, viewed 26 December 2019, https://www.fortishealthcare.com/india/fortis-escorts-hospital-in-faridabad
27. D. Dua and C. Graff., UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, viewed 17 April 2020, http://archive.ics.uci.edu/ml
28. S. Mukherjee, *Information Value (IV) And Weight Of Evidence (WOE)*, viewed 27 April 2020, https://stepupanalytics.com/information-value-iv-and-weight-of-evide-nce-woe/
29. L. McInnes., J. Healy and J. Melville *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, viewed 5 May 2020, arXiv:1802.03426
30. M. Galarny, *Understanding Boxplots*, viewed 5 May 2020, https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

AUTHORS PROFILE

Ayushi, is a research scholar currently pursuing Master’s in Computer Engineering from J. C. Bose University of Science and Technology (formerly YMCA University of Science and Technology). She has completed her Bachelor’s in Computer Science and Engineering from Lingayas GVKS Institute of Management and Technology, Faridabad affiliated to MDU Rohtak.

Shilpa Sethi, has received her Master in Computer Application from Kurukshetra University, Kurukshetra in the year 2005 and M. Tech. (CE) from MD University Rohtak in the year 2009. She has done her PhD in Computer Engineering from YMCA University of Science & Technology, Faridabad in 2018. Currently she is serving as Assistant Professor in the Department of Computer Applications at J.C. Bose University of Science & Technology, Faridabad Haryana. She has published more than thirty research papers in various International journals and conferences. Her area of research includes Internet Technologies, Web Mining, Information Retrieval System. and Artificial Intelligence.

Dr. Jyoti, PhD, is an Assistant Professor in the Department of Computer Engineering, J.C. Bose University of Science and Technology, Faridabad, Faculty of informatics and Computing. She has a total experience of 16 years with 9 years of research experience. She has authored 32 papers in reputed journals. She has guided numerous M.Tech thesis and projects at the undergraduate level.