Improving Performance of Coronary Infarction By Using Machine Learning

Sonam Rathore[1], Ankur Mudgal[2]
[1] Dept of CSE
[2]Asst. Professor, Dept of CSE
[1, 2] Shri Ram Institute of Science & Technology, Jabalpur, Madhya Pradesh, India.

Abstract- *Data mining techniques are used in various applications. In healthcare industry, it plays an important role to predict diseases. For detecting a disease, a patient had to do a number of tests, basically for heart disease which became nowadays a risk for the people as it is increasing day by day. We are living in a post modern era and there are tremendous changes happening to our daily routines which make an impact on our health positively and negatively. As a result of these changes various kind of diseases are enormously increased. Especially, heart disease has become more common these days. The life of people is at a risk. Variation in Blood pressure, sugar, pulse rate etc. can lead to cardiovascular diseases that include narrowed or blocked blood vessels. It may causes Heart failure, Aneurysm, Peripheral artery disease, Heart attack, Stroke and even sudden cardiac arrest. Many forms of heart disease can be detected or diagnosed with different medical tests by considering family medical history and other factors. But, the prediction of heart diseases without doing any medical tests is quite difficult. The aim of this thesis is to diagnose different heart diseases and to make all possible precautions to prevent at early stage itself with affordable rate. We follow 'Data mining' technique in which attributes are fed in to SVM, Random forest, KNN, and Decision Tree classification Algorithms for the prediction of heart diseases. The preliminary readings and studies obtained from this technique is used to know the possibility of detecting heart diseases at early stage and can be completely cured by proper diagnosis.*

Keywords- Data Mining, SVM, Naive Bayes, KNN, Random Forest, Heart Disease.

I. INTRODUCTION

The World Health Organization (WHO) [1] lists cardiovascular diseases as the leading cause of death globally with 17.9 million people dying every year. The risk of heart disease increases due to harmful behavior that leads to overweight and obesity, hypertension, hyperglycemias, and high cholesterol [1]. Furthermore, the American Heart Association [2] complements symptoms with weight gain (1–2 kg per day), sleep problems, leg swelling, chronic cough and high heart rate. Diagnosis is a problem for practitioners due the symptoms' nature of being common to other conditions or confused with signs of aging. The growth in medical data collection presents a new opportunity for physicians to improve patient diagnosis. In recent years, practitioners have increased their usage of computer technologies to improve decision-making support. In the health care industry, machine learning is becoming an important solution to aid the diagnosis of patients. Machine learning is an analytical tool used when a task is large and difficult to program, such as transforming medical record into knowledge, pandemic predictions, and genomic data analysis.

Recent studies have used machine learning techniques to diagnose different cardiac problems and make a prediction. Melillo et al. [3] contributed to an automatic classifier for patients with congestive heart failure (CHF) that separates patients with minimal risk from those at high risk. The classification and regression tree (CART) computed a sensitivity and specificity of 93.3% and 63.5%, respectively. Rahhal et al. [4] proposed a deep neural network (DNN) classification of electrocardiogram (ECG) signals to learn the best set of features and improved the performance. Guidi et al. [5] contributed to a clinical decision support system (CDSS) for the analysis of heart failure (HF). They compared the performance of different machine learning classifiers, such as neural network (NN), support vector machine (SVM), a system with fuzzy rules that uses CART, and random forests (RF). The CART model and RF obtained the best performance with an accuracy of 87.6%. Zhang et al. [6] found a NYHA class for HF from unstructured clinical notes using natural language processing (NLP) and the rule-based method, calculating an accuracy of 93.37%. Parthiban et al. [7] scrutinized an SVM technique to diagnose heart disease in patients with diabetes, obtaining an accuracy of 94.60% and predicting features such as age, blood pressure, and blood sugar.

A major problem of machine learning is the high dimensionality of the dataset. The analysis of many features requires a large amount of memory and leads to an overfitting, so the weighting features decrease redundant data and processing time, thus improving the performance of the algorithm [8],[9]. Finding a small set of features characterizes different diseases of health management, genome expression, medical images, and IoT. Dimensionality reduction uses feature extraction to transform and simplify data, while feature selection reduces the dataset by removing useless features.

In the literature, the use of feature selection techniques improved the prediction of heart disease. Dun et al. [10] studied the presence of heart disease through deep learning techniques, random forests, logistic regression, and SVM with hyper parameter tuning and feature selection. NN had the best accuracy at 78.3%. Sewak et al. [11] reduced cardiovascular features using the Fisher ranking method, generalized discriminant analysis (GDA), and a binary classifier as extreme learning machine (ELM). They detected coronary heart disease with an accuracy improvement of 100%. Yaghouby et al classified arrhythmias with heart rate variability (HRV). They achieved 100% accuracy using GDA for feature reduction and multilayer perceptron (MLP) neural network as a classifier. Mohammadzadeh et al. [12] classified 15 features from HRV signal. GDA reduced the features to five and computed 100% precision using SVM.

Principal component analysis (PCA) creates new components that store the most valuable information of the features by capturing a high variance. Recently, several studies have used PCA as a feature extraction technique for classification in health care. Rajagopal et al. [13] compared an automatic classification of cardiac arrhythmia using five different linear and non-linear unsupervised dimensional reduction techniques with the neural network (PNN) classifier. With a minimum of 10 components, fastICA computed an F1 score of 99.83%. Zhang et al. [14] detected breast cancer using an AdaBoost algorithm based on PCA. Negi et al. [15] combined PCA with a feature reduction technique called uncorrelated linear discriminant analysis (ULDA) to obtain the best features that control upper limb motions. Avendaño-Valencia et al. [16] applied PCA to time frequency representations (TFR) to reduce heart sounds and improve performance.

Earlier studies worked with a heart disease subset of 13 features (Subset-A). The aim of classification was to predict whether a patient had heart disease using, in most cases, the dataset of Cleveland. The classification learning models combined with dimensionality reduction seek to achieve three primary objectives: (i) to learn the best feature representation of the dataset used; (ii) to validate the performance of PCA in conjunction with a feature selection technique; and (iii) to learn the classification model that computes the best performance.

The untimely detection of heart diseases can stop the death. But in every situation it is observe at the last stages of disease. Or after death .So the Health cares are point to detect the disease at early stages. In this case Data mining technique is the good technique can detect disease. Completely cure the disease by proper diagnosis. But the main problem of Data mining is using different algorithms for detection of heart disease. Some algorithms are diagnose is less accurate and time consuming. Next problem is taking the survey less number of attributes is used in previous papers and also less number of patients is used.

The main motivation of this research is to provide an accurate disease diagnosis framework with reduced feature set. Manually, doctors have to perform number of tests in order to diagnose a particular disease which requires a lot of time, effort and money. Automated disease diagnosis system will predict the heart disease with high accuracy resulting in time and effort reduction.

II. RELATED WORK

Liaqat Ali et.al. [17] Introduce an expert system that stacks two support vector machine (SVM) models. The first SVM model is linear and L1 regularized. It has the capability to eliminate irrelevant features by shrinking their coefficients to zero. While the second SVM model is L2 regularized. It is used as a predictive model. To optimize the two models, we propose to use a hybrid grid search algorithm (HGSA) which is capable of optimizing the two models simultaneously. The effectiveness of the proposed method is evaluated using six different evaluation metrics including accuracy, sensitivity, specificity, MCC, ROC charts and area under the curve (AUC). The first model has the capability to eliminate irrelevant features by shrinking their coefficients to zero. Performance comparison is done using accuracy, ROC chart and AUC evaluation metrics. The proposed method is efficient in terms of time complexity.

Ashir Javeed et. al. [18] develops a novel diagnostic system. The proposed diagnostic system uses random search algorithm (RSA) for features selection and random forest model for heart failure prediction. The proposed diagnostic system is optimized using grid search algorithm. Two types of experiments are performed to evaluate the precision of the proposed method. In the first experiment, only random forest model is developed while in the second experiment the proposed RSA based random forest model is developed. Experiments are performed using an online heart failure database namely Cleveland dataset. In addition, the proposed method achieved classification accuracy of 93.33% while improving the training accuracy as well. We develop only random forest model which is implemented in Python programming package. The model hyper parameters are tuned using grid search algorithm. It was shown that the proposed RSA-RF learning system improves the performance of random forest model by 3.3%. It was also observed that the proposed system reduces the time complexity of the machine learning models by reducing the number of features.

Authors have proposed different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. [19]

After data munging and attributes selection, machine learning algorithms including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine(SVM) and Adaptive Boosting, are used for prediction of the above-mentioned diseases, and a comparison of their accuracy is done for selecting best model for that disease dataset. The feature selection from 13 input parameter by backward elimination resulted in a total of 11 significant input parameters which include gender, type of chest pain, blood pressure, blood sugar level, electrocardiograph result, maximum heart rate, exercise induced angina, old peak, Slope, number of vessels colored, thal.

Logistic Regression was found to have the highest accuracy among all. The prediction accuracy of our proposed method reaches 87.1% in Heart Disease detection using Logistic Regression.

Anjan Nikhil Repaka et. al. [20] Smart Heart Disease Prediction) is built via Navies Bayesian in order to predict risk factors concerning heart disease. For predicting the chances of heart disease in a patient, the following attributes are being fetched from the medical profiles, these include: age, BP, cholesterol, sex, blood sugar etc... The collected attributes acts as input for the Navies Bayesian classification for predicting heart disease. This classification algorithm basically employs conditional independence; this implies that value of an attribute for an available class is not dependent on other attribute values since the algorithm relies upon the Bayesian theorem. Proposed classification techniques performance which is compared with prevailing techniques of SMO (Sequential Minimal Optimization), Bayes Net and MLP (Multi-Layer Perception). Effective outcome is exhibited by the proposed Navies Bayesian with greater performance in contrast to rest of the techniques.

Amogh Powar et. al. [21] Data mining gears are expecting destiny traits therefore permitting making understanding driven selections. There are many statistics mining techniques like Decision Tree, Naïve Bayes, & Neural Network. This paper proposes to use three of those techniques for predicting the coronary heart ailment. Each of these techniques will be tested in the various parameters and different accuracies will be obtained with the different parameters. In this exploration we have endeavoured to overview twenty research papers, which fundamentally talks about the different normal information mining and computerized reasoning systems that can be utilized for forecast of heart maladies and shows which technique is the most precise.

III. PROPOSED WORK

Figure 1 represents the overall architecture of the proposed system. First step is to collect dataset; one of the sources for it is Cleveland Clinic Foundation. After this data pre-processing is performed, this includes cleaning of data, dealing with missing values & its features. After pre-processing the dataset, it is split into training set and testing set. Training set is used to train the algorithm and testing set is used for testing purpose. Proposed algorithm takes training dataset as the input to train on various samples and produces a trained model based on proposed algorithm. Proposed algorithm is random forest. Testing data is than applied on the model to predict the result.

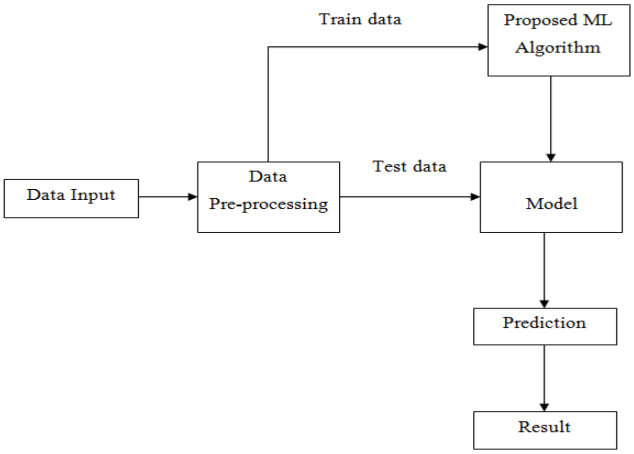

Fig 1: Proposed architecture

Proposed system is developed with a machine learning approach known as random forest. Random Forest, also known random decision forest is associated with the predictive models for both the classification and regression problems. Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. In random forest, the models create an entire forest of random uncorrelated decision and are trained with bagging method. Bagging method is a combination of learning models to increase the overall result.

The random forest technique is a type of additive model that predicts the data by combining decisions from a sequence of base models. It reduces the variance by avoiding over fitting of the model. Random forest arrives at a decision or prediction based on the maximum number of votes received from the decision trees. The outcome which is arrived at, for a maximum number of times through the numerous decision trees is considered as the final outcome by the random forest. Features are selected randomly using a method known as bootstrap aggregating or bagging. From the set of features available in the dataset, a number of training subsets are created by choosing random features with replacement. For example, if a dataset contains 20 features and subsets of 5 features are to be selected to construct different decision trees then these 5 features will be selected randomly and any feature can be a part of more than one subset. This ensures randomness, making the correlation between the trees less, thus overcoming the problem of over fitting. Once the features are selected, the trees are constructed based on the best split. Each tree gives an output which is considered as a 'vote' from that tree to the given output. The output which receives the maximum 'votes' is chosen by the random forest as the final output/result or in case of continuous variables, the average of all the outputs is considered as the final output.
Proposed algorithm is as follows:

1. We first take a random sample of size N with replacement from the data.
2. Take a random sample without replacement of the predictors.
3. Construct a split by using predictors selected in step 2.
4. Repeat steps 2 and 3 for each subsequent split until the tree is as large as desired.
5. Drop the out-of-bag data down the tree. We then store the class assigned to each observation along with each observation's predictor values.
6. Repeat steps 1–5 for large number of times.
7. For each observation in the dataset, we count the number of trees that it is classified in one category over the number of trees.
8. Assign each observation to a final category by a majority vote over the set of trees. Thus, if 51% of the time over a large number of trees a given observation is classified as a "1" that becomes its classification

IV. RESULT & EVALUATION

The database for this research work has been taken from the StatLog dataset in UCI repository. It includes 13 attributes. The heart disease dataset included in this research work consists of total 270 instances with no missing values. The dataset is typically used for various types of heart diseases such as typical angina, atypical angina, and non-anginal pain and asymptomatic. The aim of this research is to know whether the patient has heart disease or not. The records in the datasets are divided into training set and test sets. After preprocessing the data, data mining classification technique namely decision tree, naïve Bayes, KNN, SVM and Random Forest were applied. This section shows the results of those classification model done using Python Programming. The results are generated for both training datasets and test data sets. Following pictures are the main outcome of data mining technique is that we can predict the heart diseases and can receive precautions. The classification performance can be evaluated in three terms: accuracy as defined below. Accuracy explains correctly classified instances of the symptoms with respect to heart disease. Fig below represents accuracy achieved by proposed method.

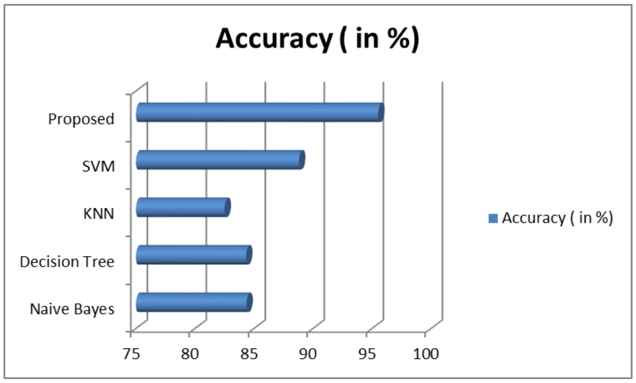

Fig 2: Accuracy chart

V. CONCLUSION

The main motivation of this thesis is to provide an insight about detecting and curing heart disease using data mining technique. For this thesis, data were collected from Kaggle Data Sets. All attributes are numeric-valued. The data was collected from the four locations. It is integer valued from 0 (no presence) to 4 to predict the likelihood of patient getting heart diseases. These attributes are fed in to Naive Bayes, SVM, KNN, Decision Tree and Random forest, in which Random Forest gave the best result with the highest accuracy. Valid performance is achieved using Random Forest algorithm in diagnosing heart diseases and can be further improved by increasing the number of attributes. Thus, in an environment similar to that of the used dataset, if all the features are preprocessed such that they acquire normal distribution, Random Forest is a good selection to obtain a robust prediction model. And, such models provide a valuable assistant to the society for health care management domain.

REFERENCES

[1] Cardiovascular diseases (CVDs) retrieved from http://www.who.int/cardiovascular_diseases/en/ (2019, July 16) Google Scholar.

[2] American Heart Association Classes of heart failure Retrieved from https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure (2018, August 11), Google Scholar.

[3] P. Melillo, N.D. Luca, M. Bracale, L. Pecchia "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability", IEEE J. Biomed Health Inf, 17 (3) (2013), pp. 727-733, 10.1109/jbhi.2013.2244902 View Record in ScopusGoogle Scholar.

[4] M.A. Rahhal, Y. Bazi, H. Alhichri, N. Alajlan, F. Melgani , R. Yager, "Deep learning approach for active classification of electrocardiogram signals Inf Sci, 345 (2016), pp. 340-354, 10.1016/j.ins.2016.01.082 Article Download PDF View Record in ScopusGoogle Scholar.

[5] G. Guidi, M.C. Pettenati, P. Melillo, E. Iadanza , "A machine learning system to improve heart failure patient Assistance IEEE J Biomed Health Inf., 18 (6) (2014), pp. 1750-1756, 10.1109/jbhi.2014.2337752 Cross Ref View Record in ScopusGoogle Scholar

[6] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, S. Spe edie Automatic methods to extract New York heart association classification from clinical notes IEEE Int Conf Bioinformatics Biomed (BIBM) (2017), 10.1109/bibm.2017.8217848, 2017.

[7] G. Parthiban, S.K. Srivastava Applying machine learning methods in diagnosing heart disease for diabetic patients Int J Appl Inf Syst, 3 (7) (2012), pp. 25-30, 10.5120/ijais12-450593.

[8] M. Yang, Y. Nataliani, "A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy", IEEE Trans Fuzzy Syst, 26 (2) (2018), pp. 817-835, 10.1109/tfuzz.2017.2692203.

[9] R. Chen, N. Sun, X. Chen, M. Yang, Q. Wu "Supervised feature selection with a stratified feature weighting method", IEEE Access, 6 (2018), pp. 15087-15098, 10.1109/ACCESS.2018.2815606.

[10] B. Dun, E. Wang, S. Majumder Heart disease diagnosis on medical data using ensemble learning (2016).

[11] R.S. Singh, B.S. Saini, R.K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine", Clujul Med, 91 (2) (2018), p. 166, 10.15386/cjmed-882.

[12] B.M. Asl, S.K. Setarehdan, M. Mohebbi, "Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal ", Artif Intell Med, 44 (1) (2008), pp. 51-64, 10.1016/j.artmed.2008.04.007.

[13] R. Rajagopal, V. Ranganathan, **"**Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification", Biomed Signal Process Contr, 34 (2017), pp. 1-8, 10.1016/j.bspc.2016.12.017

[14] D. Zhang, L. Zou, X. Zhou, F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer", IEEE Access, 6 (2018), pp. 28936-28944, 10.1109/access.2018.2837654.

[15] S. Negi, Y. Kumar, V.M. Mishra, "Feature extraction and classification for EMG signals using linear discriminant analysis"**,** 2016 2nd international conference on advances in computing, communication, & automation (ICACCA) (2016), 10.1109/icaccaf.2016.7748960.

[16] D. Avendano-Valencia, F. Martinez Tabares , D. Acosta-Medina, I. Godino G. Castellanos-Dominguez **"**TFR-based feature extraction using PCA approaches for discrimination of heart murmurs", 2009 annual international conference of the IEEE engineering in medicine and biology society (2009), 10.1109/iembs.2009.5333772.

[17] Liaqat Ali, Awais Niamat, Javed Ali Khan, Noorbakhsh Amiri Golilarz, And Xiong Xingzhong," An Expert System Based on Optimized Stacked Support Vector Machines for Effective Diagnosis of Heart Disease', 2169-3536 (c) 2018 IEEE.

[18] Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor, Redhwan Nour, Samad Wali and Abdul Basit," An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection", DOI 10.1109/ACCESS.2019.2952107, IEEE Access.

[19] Pahulpreet Singh Kohli, Shriya Arora," Application of Machine Learning in Disease Prediction", 2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE.

[20] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin," Design And Implementing Heart Disease Prediction Using Naives Bayesian" , Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8.

[21] Amogh Powar, Seema Shilvant, Varsha Pawar,” Data Mining &Artificial Intelligence Techniques for Prediction of Heart Disorders: A Survey”, 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), IEEE-2019.