iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33848

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Review on Heart Disease Prediction using Machine Learning

Prateek Sharma[1], Naveen Tiwari[2], Ayushmaan Verma[3], Saahil Yadav[4], Paranjay Bhatt[5], Prashashti Kanikar[6]
[1, 2, 3, 4]*Computer Engineering, Mukesh Patel School of Technology Management and Engineering, NMIMS, India*

Abstract: ***According to World Health Organization (WHO) reports, the heart disease(s) are the number one cause of death globally. More people die annually from heart diseases than from any other cause. An estimated 17.5 million people died from these diseases in a single year, 2017, representing 31% of all global deaths. Of these deaths, an estimated 7.4 million were due to coronary heart disease. In order to decrease mortality from heart diseases there should be a fast and effective detection method. Machine learning can be a convenient tool to assist doctors in predicting the disease by obtaining knowledge and information regarding the disease from past patient's data.***

Keywords: Machine Learning, Naïve Bayes, Decision Tree, Random Forest, K-NN, Bagging, Boosting, Logistic Regression, Stacking, Ensembled Learning, Hard Voting Classifier, Multilayer Perceptron, Hyperparameter tuning, Neural Network, SGD Classifier, etc.

I. INTRODUCTION

Heart Disease depicts a scope of conditions that influence a person's heart. 'Heart disease' term incorporates various diseases, for example, coronary artery disease; heart rhythm problems (arrhythmias); and heart defects that a person is born with, better known as congenital heart defects, among a lot of others. Heart disease can sometimes also be referred to as cardiovascular disease. Cardiovascular disease (CVD) usually refers to a condition in human heart that narrows down or block blood vessels that can lead to a heart attack (Myocardial infarctions), chest pain or stroke. Various other heart conditions, for example, the ones that influence our heart's muscle, valves, or rhythms, additionally are viewed as types of heart disease. On an average 17.9 million people die each year from CVDs, which is an estimated 31% [16] of all deaths globally. In today's world, health care area delivers an enormous measure of data about patients, disease determination, and so forth anyway this information is not utilized effectively by the professionals and researchers. Healthcare industry often suffers when it comes to providing quality of service in many regions of the world. It means diagnosing the disease correctly & providing immediate and effective treatments to patients. Poor diagnosis can be a cause to severe consequences which can be extremely harmful and is totally unacceptable. There are various risk factors that can leads to this disease such as family history, increasing age, etc. are some risk factors that cannot be controlled. But unhealthy lifestyle, smoking habits, high blood pressure, sedentary lifestyle, problems like overweight or obesity comes under those factors that can be controlled or prevented in different ways.

With the advancement in technology, we need a more reliable and economical method to help people detect and diagnose the disease that takes the greatest number of lives worldwide. Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

[17] Machine learning (ML), a branch of artificial intelligence (AI) that has been increasingly utilized in healthcare industry and medicine recently. It is basically how computers understand and make sense of data and decide the outcome or classify a task with or without human supervision. The conceptual framework of ML is based on models that receive input data in form of records that are obtained from patients and the outcome is predicted by combining mathematical optimization and statistical data analysis. Selection of optimal algorithms for application in different clinical datasets may seem feasible, but the clinical interpretation and decision of implementing algorithms can be very challenging. A profound understanding and thorough knowledge of statistical and clinical knowledge in developers is additionally a challenge.

II. PREREQUISITE

A. *Classification* [17]

Classification is a supervised machine learning technique It may be defined as the process of predicting class or category from observed values or given data points. To implement classification, we first need to train the classifier. This is done by splitting the dataset into a training set and testing set. After successfully training the classifier, it can be used on the testing set to check for the accuracy.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

B. Naïve Bayes [17]

A Naive Bayes classifier is a probabilistic machine learning model which is can be used to deal with classification problems. This classifier algorithm is based on the Bayes formula.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes theorem can be used to find the probability of A happening, given that B has occurred, where **B** is the evidence and **A** is the hypothesis. We assume here that the predictors/features are independent which signifies that the presence of one particular feature does not affect the other. Hence it is called naive.

C. Decision Trees [17]

Decision Tree, a supervised learning method, is the classifier that has the structure of a tree in which the features of the dataset are represented by internal nodes, branches represent the decision rules, and the outcome is represented by each leaf node. This algorithm can be used for both classification and regression problems, however it mostly used for solving classification problems. This algorithm is basically a graphical representation for acquiring all the possible results to a problem based on the given conditions and make a best decision for the same.

D. Hyperparameter Tuning[17]

This is the process of tuning that determines the optimal value for the attributes, and it increases the accuracy of the model. Basically, it finds the correct value for each parameter at which the model gives the highest possible accuracy. It can be executed using two methods: GridSearchCV and RandomizedSearchCV.

E. Hard Voting Classifier[18]

This is the ensemble technique which combines the different classifiers, and it predicts the result by simple majority vote to compute the accuracy. Due to various combination of strong classifiers, it gives very high accuracy as compared to other ensemble techniques like random forest, gradient boosting etc.

F. Random Forest[17]

Random forest is a supervised learning algorithm that builds multiple decision trees and merges them together randomly to get a more accurate and stable prediction. The forest built in this algorithm, is essentially an ensemble of decision trees, that are trained and votes from different decision trees is aggregated to decide the final class of the test object.

G. KNN[17]

K-Nearest Neighbours is a supervised machine learning algorithm that can be used to deal with both classification and regression problems.

It assumes same data points in the close proximity. Basically, whenever the new data point arrives between the 2 classes the class having majority of the data points nearer to that new data point than this new data point belongs to that class. Some rules are to be followed before selecting the value of K and usually the K value is taken as odd to avoid confusion.

H. Logistic Regression[17]

Logistic regression is a statistical model which is used to determine if an independent variable has an effect on dependent variable which is a binary variable. It uses a sigmoid function, so it squeezes the output value between 0 and 1. There is also another reason why it formed S shaped curve because it predicts the probability which is between 0 and 1 only that is why it forms an S shaped curve.

I. Neural Network[18]

A neural network is a series of algorithms that essentially mimics the operating of a human brain by pointing and recognizing the relationships in a set of data. A neural network consists of units called neurons which are arranged in layers that are defined to be feed forward, i.e., a unit feeds its output as an input to all the units in the next layer. All the units pass signals to the next unit on which the weightings are applied. These weightings are used to help the network learn and adapt by tuning them in the training phase.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

III. LITERATURE SURVEY

In 2012, T.John Peter and K. Somasundaram[1] used different machine learning algorithms like Decision Tree, Naïve Bayes, K-NN and Neural network to predict heart disease. They proposed a technique to create to Attribute Relation File Format (ARFF) file which contains list of attributes and it is an ASCII file. Usually, datasets are large, and classification takes long time to predict disease but, in this study, they have used different dimensionality reduction technique on dataset and then they are classified using Decision Tree, Naïve Bayes, K-NN and Neural network. Naïve Bayes gives maximum accuracy of 83.7% and after using CFS subset for dimensionality reduction its accuracy increases by 1.8%.

In 2014, Muhammad Fathurachman, Umi Kalsum, Noviyanti Safitri and Chandra Prasetyo Utomo[2] used extreme learning based neural networks for heart disease prediction and compared the results obtained from applying other classification algorithms namely Decision Trees (DT), Back Propagation Artificial Neural Networks (BP-ANN), and Support Vector Machine (SVM). 5-fold cross validation is applied on the data and the results obtained are compared. The data set used records 270 instances with 14 different attributes. This dataset is split into training set and testing set in the ratio of 60:40. The results are compared on the basis of 5-fold cross validation. Based on the comparison of these results, it is concluded that the Extreme Learning Method (ELM) is the best performance algorithm with an accuracy of 83%.

In 2016, SENTHILKUMAR MOHAN, CHANDRASEGAR THIRUMALAI and GAUTAM SRIVASTAVA[3] introduced a new technique called Hybrid Random Forest with Linear Model (HRFLM) its main objective was just to give better accuracy as compared to the existing models. This algorithm even gives better results than the neural network. For this study they have used Cleveland UCI repository dataset having 303 instances among them 6 were missing values so they were removed during pre-processing and have 14 attributes including target variable. They have also used other existing algorithms to compare the accuracy and to proof their model is best till now. So, as a result HRFLM gives accuracy of 88.4%, precision of 90.1%, sensitivity of 92.8%, specificity of 82.6% and F1 score of 90% which is highest among all the used classifiers. Hence this Hybrid Random Forest with Linear Model gives best results. The UCI dataset later on divided into 8 datasets and all were classified using R Studio Rattle. It was observed that in data split 4 random forest have the highest error rate of 20.9% and in data split8 linear model have lowest error rate of 9.1% as compared to random forest and decision tree classifier. So random forest was combined with the linear model in order to improve the accuracy.

In 2018, ASHIYA ZABEEN, ANKUR UTSAV and KANHAIYA LAL [4] applied fuzzy skilled system to predict the cardiovascular disease and the main objective of this study is to detect disease at an early stage with improved accuracy as compared to the neural network. For this study, they have used dataset from UCI [19] repository having 7 input attributes and 1 target variable. First of all, pre-processing is done and in data analysis 7 important attributes (age, Blood pressure, Cholesterol, E.C.G, Blood sugar, Heart rate, Thallium scan.) were identified. After this step fuzzification is done in which inputs and outputs were determined that they belong to which corresponding fuzzy set. Membership function for each fuzzy set in each attribute is classified as trapezoidal or triangular. After these 50 different fuzzy rules were identified after the analysis of datasets and after this defuzzification was done predict the results. It is observed that this proposed system has accuracy of 80% more than the neural network.

In 2018, Ming Liu and Younghoon Kim [5] proposed a classification technique for heart diseases dependent on ECG using Long Short-Term Memory (LSTM), a very effective deep learning method that analyses time series very efficiently. The data used by the authors here is obtained from archives of UCR Time Series. This data is pre-processed in two steps: extracting heartbeats and normalizing the heartbeats' lengths using interpolation. A quick and efficient method called as Symbolic Aggregate Approximation (SAX) is used as a pre-processing method for extensive evaluation to check and improve the accuracy of the model. 4 different classification methods are evaluated and compared, which are LSTM, Move-Split-Merge (MSM), Dynamic time Warping (DTW), and Complexity Invariant Distance (CID). Furthermore, varying hyper-parameters are used as well and the best SAX parameters that influences the accuracy in each model are extracted. The paper concludes that using LSTM with pre-processing method SAX gives an accuracy of 97%.

In 2018, Shriya Arora and Pahulpreet Singh Kohli[6] three different classification algorithms were applied on three different datasets that had data for three separate diseases (Diabetes, Breast cancer and Heart Disease). The proposed methodology was proceeded as: Exploration of data, Data munging, feature selection and execution of the algorithms. All the before mentioned 3 algorithms were applied on the heart dataset that had 13 attributes and the observations were recorded. The results show that the best algorithm that can be applied to the heart dataset to give accurate results is Logistic Regression which gives an accuracy of 87.1%.

In 2019, Rahma Atallah and Amjed Al-Mousa[7] applied Machine Learning algorithms such as Stochastic Gradient Descent (SGD) Classifier, K-Nearest Neighbor Classifier, Random Forest Classifier and Logistic Regression Classifier and these 4 algorithms were combined using Hard Voting ensemble method and the result is predicted based on the majority votes from all the models.

This ensemble method is used in order to improve the accuracy and predict heart disease with very low cost. The Heart Disease Dataset is used from Cleveland UCI repository which consists of 303 instances and 14 attributes. The attributes are as follows: age, chest pain, sex, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic, maximum heart rate, ST depression, peak exercise slope, number of major vessels colored by fluoroscopy, heart rate and target. So, in pre-processing Min-Max scaler is used as a data normalization technique and all data is converted into range of 0 and 1. Stochastic Gradient Classifier give the highest accuracy of 88% as compared to KNN, Random forest and Logistic Regression which give the same accuracy of 87%. Hyperparameter tunning is done using Grid SearchCV to find the optimum value of the parameters instead of using the default values of the parameters which in turn increases the accuracy for each algorithm. The hard voting ensemble method is used to combine these 4 algorithms to improve the accuracy and the result is predicted based on the majority votes of the model. Hence Hard Voting Ensemble gives highest accuracy of 90% as compared to the other algorithms.

In 2019, Amin Ul Haq, Jianping Li[8] in this study develop a prediction system by implementing ensemble learning techniques using Feature Selection Relief algorithms along with different machine learning classification algorithms for heart disease prediction. The proposed consists of these major parts: pre-processing of attributes, features selection based on relief, Machine learning classifiers, applying cross-validation methods, and performance evaluation. The dataset taken is the one from UCI repository's Cleveland database with 297 records and 13 essential attributes. The algorithms used are Logistic Regression (LR), Support Vector Machine (SVM- RBF), K-Nearest-Neighbour (K-NN), Decision Tree (DT), Naive Bayes (NB), and Artificial-Neural-Network (ANN). Further ensembled learning methods are used to define new algorithms using bagging, boosting and stacking. Bagging is similar to voting as it is necessarily the integration of the predictions. The only difference in Boosting is that the efficiency obtained in previous models affect the new models. When we combine models of various types, it is known as Stacking. Additionally, to improve the overall results of the model, K-fold cross validation methods have been applied. (K=10). The results obtained when the model is trained based on the algorithms before applying ensembled learning methods show that SVM- RBF gives the most accurate results with an accuracy of 86%.

The different predictive models were then combined (ensembled) using the stacking method with increased accuracy of the classifiers. The results of different combination show that an ensembled model of ANN, Logistic regression, and SVM(RBF) gives the best accuracy.

Furthermore, we get to see in this paper that a model for heart disease prediction is also trained with Backward Propagation Neural Network (BPNN) algorithm. 3 networks are trained using the same activation function, RelU and different parameters to get a result with 93% best accuracy.

In 2019 ASHIR JAVEED, SHIJIE ZHOU, LIAO YONGJIAN, IQBAL QASIM, ADEEB NOOR and REDHWAN NOUR[9] proposed a hybrid method which combines Random Search Algorithm (RSA) and Random Forest (RF) which gives very high accuracy of 93.3%. This study is divided into three sections: in first section random forest is implemented using Python programming and hyperparameter tuning is done using grid search method to optimize the results and accuracy achieved in this is 90%; in second section Random Search Algorithm is used for feature selection which computes different subsets of features from 1 to N-1 range and these features are used to train the model using random forest and hyperparameter tuning is also done to optimize the results and due to this 93.3% accuracy is achieved; in third section the proposed RSA-RF method is compared with Adaboost ensemble, extra tree ensemble, Random Forest, Linear SVM, SVM(RBF) and as a result the proposed hybrid method gives highest accuracy of 93.3% at features equals to 7 which is calculated by RSA.

In 2019, Anjali Kumari, Divya Krishnani, Akash Dewangan[10] and others analysed and estimated how various machine learning algorithms can be used for the prediction of heart disease in their paper. The dataset used by the authors is a part of Framingham Heart Study dataset which records 16 attributes for 4240 participants including attributes like diabetes, gender, age, BPMeds, BMI, etc. The writers applied different classification algorithms such as Random Forest, Decision Tree and K-Nearest Neighbours classification models. K cross-validation is also applied for all the algorithms used where k is taken to be 10. The results obtained from all the three classification algorithms are obtained and compared. Similarly, the accuracy obtained after each fold of cross validation is compared and it is concluded that as the k-fold increases, the accuracy increases as well. The obtained results' performance evaluation is done based on the measures such as Accuracy, Recall/Sensitivity, Specificity, Precision and F1 Score. As per the observations, the authors observed that Random Forest gives the best accuracy.

In 2019, Amin Ul Haq and Jianping Li[11] authored another research paper based on a heart disease prediction system using Feature Selection. In this paper, they used the Cleveland heart disease dataset that has 13 attributes and records data of 303 participants. The dataset was split in 70:30 training to testing set ratio. After the data is pre-processed, a classical feature selection method, known as Sequential Backward Selection (SBS) algorithm is applied.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

This method is meant to extract the features with minimum total round trip time in classifier performance, hence reducing the model execution time and can improvement in prediction of the model can be improved. After the feature selection is processed, K-Nearest Neighbour (K-NN) algorithm is applied to build a predictive model. For each number of features extracted, K-nearest neighbours was used. In this paper, 8 different values of k (1, 2, 3, 4, 5, 6, 7, 8) are used for building the predictive model and the results obtained at every observation is compared and, in the end, the mean value of all these results is calculated to get the final accuracy. It is observed that the different number of features when calculated at different k values give different accuracies. The most accuracy is obtained when only 6 features are extracted.

In 2020, Sachin Singh, Animesh, and Thomas Penzel[12] applied neural networks to classify ECG samples into three types of disorders namely arrhythmia, normal sinus rhythm, and congestive heart failure. There were three different datasets used collectively to build this model. A total of 162 patients' readings (65536 samples in total) whose samples were recorded in the datasets consisting of the three aforementioned classes. The main objective of the authors here was to train a classifier for differentiating between arrhythmia (ARR), normal sinus rhythm (NSR), and congestive heart failure (CHF). The proposed methodology has been analysed using two different approach as follows:

A. Methodology 1 Process Flow

1) Input Data from PhysioNet
2) Over Sampling
3) Pearson Correlational Analysis
4) Selection of 0.5 percentile best features
5) ANOVA
6) Feature Extraction

In this methodology, after the feature extraction algorithms were applied, Random Forest and Support Vector Machine algorithms are used to train the classifier model where Random Forest gives better accuracy than SVM.

B. Methodology 2 Process Flow

7) Input Data from PhysioNet
8) Over Sampling
9) Multivariate autoregressive (AR) modelling
10) ANOVA
11) Feature Classifier

After the feature selection and feature extraction algorithms were applied, this methodology is approached by application of different algorithms namely: Random Forest, Support Vector Machine, LSTM and Bi-directional LSTM. The results observed shows that accuracy obtained is maximum in Bi LSTM (4 hidden layers)

Hence, the study concluded that using neural network algorithms to classify arrhythmia, normal sinus rhythm, and congestive heart failure can be done by different approaches and after applying different algorithms for feature selection, feature extraction and training the model, Bi-directional LSTM gives the best accuracy i.e., 99.12 %.

In 2020, Samir S Yadav, Shivajirao M. Jadhav & Snigdha Nagrale[13] published an article which was an application of ML for the Prediction of Heart Disease to apply different classification algorithms on a dataset for thorough analysis of various algorithms in machine learning and develop a better model with an algorithm of better accuracy. The data set used is the Cleveland dataset of cardiac disease which contains 303 records with 76 attributes. The proposed model entails the methods of Dataset Description, Classification, Comparison of methods of classification, cross-validation and performance evaluation. The classification algorithms used in this published article are: Logistic Regression, K-means clustering, K-Nearest Neighbour, Naïve Bayes, Neural Network, Fuzzy K- Nearest Neighbours and K Means Clustering with Naive Bayes Classifier. The problem of overfitting in neural networks have also been attended to by applying regularisation. It is used to regularise the increasing weights of the target function and also avoids overfitting. A number of evaluation tests were carried out to evaluate the performance of classifiers. These parameters were TP (true positive); TN (truly negative); FP (False Positive); FN (False Negative)

These parameters are used to define the confusion matrix for all the algorithms applied. Hence, it is observed that when neural network is applied on the dataset to create a model, it gives an accuracy of 98%.

In 2020, Meenu Bhatia and Dr. Dilip Motwani[14] applied machine algorithms such as bagging, boosting, majority voting and stacking these are the ensemble techniques which were used by them to improve accuracy by combining the weak learners together. The main objective of this study is to develop a system which comprises of two modules: Doctor Login and Patient Login and to predict the coronary disease with less features and with very low cost. The dataset is used from the Kaggle it consists of 8 traits in which 6 are the downright traits and 2 are numeric characteristics. Attributes are as follows: ID, age, cholesterol, venous pressure, height, weight. Patient login module of the system provides rights only to registered patient to see his/her case history and Doctor login module provides rights to assigned doctor to generate patient's case history, medications and to update patient's details.

It has been observed that majority voting classifier gives the highest accuracy (81.82%) as compared to bagging, boosting and stacking. Stacking also gives quite satisfying accuracy. The proposed ensemble technique is compared with the other classifiers such as Naïve Bayes, Decision Tree and SVM on the basis of Accuracy, Sensitivity and Specificity. As a result, proposed ensemble techniques give better results and it is also noted that lion's share casting a ballot while ensemble the feeble classifiers improves the precision by 7.26% which was the highest.

In 2020, Mrs. Kelibone Eva Mamabolo and Dr. Moeketsi Mosia[15] demonstrated that in order to identify best classifier the result should not depend totally on overall accuracy because sometimes classifier having best overall accuracy may have less accuracy of a particular class which is predicted more efficiently by a less overall accurate classifier. So, in this study dataset used is from Cleveland UCI repository which consist of 14 attributes and 303 instances. In this study Decision Tree, Logistic Regression, Multilayer perceptron and Naïve Bayes are used as a classifier to predict disease with the help of WEKA tool. It is observed that Logistic Regression gives the maximum overall accuracy of 57% as compared to other algorithms and in this study, dataset is divided into 4 classes each representing the different sick peoples. It is demonstrated that Decision Tree have less overall accuracy (53%) as compared to the Naïve Bayes (55%) but it has predicted better results for class 0 as compared to Naïve Bayes. So, it is noted that despite overall accuracy individual class accuracy also matters and classifier having less overall accuracy can give better accuracy for particular class as compared to the classifier giving higher overall accuracy.

IV. COMPARATIVE ANALYSIS OF LITERATURE SURVEY

TABLE I
Comparison Table For The Papers

Authors and Year	Techniques Used	Findings	Accuracy
T.John Peter,K. Somasundaram(2012)	Naïve Bayes, Decision Tree, Neural Network, KNN, Attribute Relation File Format (ARFF) creation	Dimensionality reduction techniques are used to predict results faster. CFS is used on Naïve Bayes which increase its accuracy.	Naive Bayes (85.5%)
Muhammad Fathurachman, Umi Kalsum, Noviyanti Safitr, Chandra Prasetyo Utomo (2014)	Extreme learning based neural network (ELM), Decision Tree, SVM, Back Propagation ANN	Extreme learning method along with hyperparameter tuning gives great accuracy.	ELM (83%)
SENTHILKUMAR MOHAN, CHANDRASEGAR THIRUMALAI, GAUTAM SRIVASTAVA (2016)	Hybrid Random Forest with Linear Model (HRFLM)	HRFLM compared with other existing models as a result it gives maximum accuracy.	HRFLM (88.4%)
Shriya Arora, Pahulpreet Singh Kohli (2018)	Logistic Regression, SVM, Adaptive Boosting, Decision Tree, Random Forest	Used 3 different disease (heart, breast cancer, diabetes) dataset to predict result for each disease and for feature selection in each dataset backward modelling using p-value test is used.	For heart disease it gives highest accuracy of 87.1%

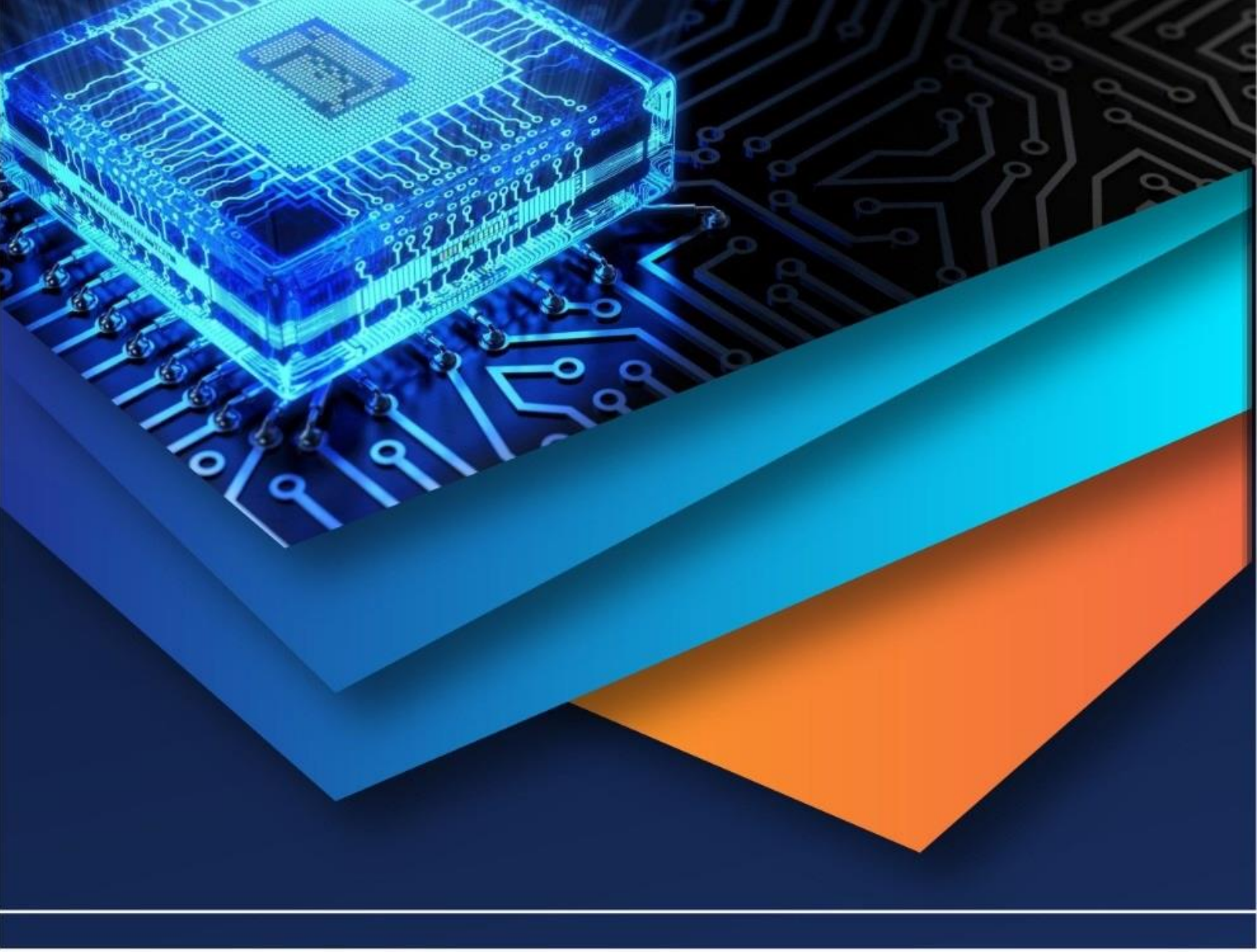
Ming Liu, Younghoon Kim (2018)	Long Short-Term Memory (LSTM), Dynamic time traveling (DTW), Move-split-blend (MSM), Complexity invariant distance (CID)	ECG computes accuracy very quickly and in this study SAX and FFT are used as pre-processing methods to achieve higher accuracy.	LSTM gives highest accuracy of 97%
ASHIYA ZABEEN, ANKUR UTSAV, KANHAIYA LAL (2018)	Fuzzy skilled system	Fuzzy logic gives better results as compared to the neural network.	Fuzzy system gives accuracy of around 80%.
Amin Ul Haq, Jianping Li (2019)	KNN	When 6 features are selected using sequential backward selection algorithm, KNN gives the most accuracy for different values of K.	With number of features equals to 6 it gives highest accuracy of 90%.
ASHIR JAVEED, SHIJIE ZHOU, LIAO YONGJIAN, IQBAL QASIM, ADEEB NOOR, REDHWAN NOUR (2019)	Random Forest, RSA-RF (Random Search Algorithm and Random Forest)	RSA is used as a feature selection technique and combined with random forest and as a result this proposed techniques gives 3.3% higher accuracy than the random forest.	RSA-RF (93.3%)
Divya Krishnani, Anjali Kumari, Akash Dewangan (2019)	Random Forest, Decision Tree, KNN	Here these three algorithms are used along with hyperparameter tuning (K =10) to give random forest as the most accurate model.	Random Forest (96.71%)
Amin Ul Haq, Jianping Li (2019)	Ensemble (ANN, Logistic Regression, SVM-RBF), Backward Propagation Neural Network (BPNN),KNN, Decision Tree, Naïve Bayes.	Predictive model using ensemble learning method is developed and is compared with model developed with neural network.	Ensemble (92%), BPNN (83%)
Rahma Atallah, Amjed Al-Mousa (2019)	Stochastic Gradient Descent Classifier, K-Nearest Neighbour Classifier, Random Forest Classifier, Logistic Regression Classifier, Hard Voting Classifier	Hard Voting classifier combines 4 algorithms and gives final accuracy by majority votes.	Hard Voting Classifier (90%)
Samir S Yadav, Shivajirao M. Jadhav & Snigdha Nagrale (2020)	Neural Network, Naïve Bayes, Logistic Regression, K-means clustering, Fuzzy KNN.	Neural Networks when regularized gives better accuracy as compared with other existing models.	Neural Network (98%)
Sachin Singh, Animesh,Thomas Penzel (2020)	Methodology1: Random Forest, SVM Methodology2: Bidirectional LSTM, Random Forest, SVM, LSTM	Two methodologies were proposed wherein methodology 1 gives random forest as most accurate model after feature extraction and methodology 2 shows Bi-directional LSTM gives excellent accuracy after feature extraction.	Random Forest (96.92%), Bi LSTM (99.12%)
Meenu Bhatia, Dr. Dilip Motwani (2020)	Bagging, Boosting, Majority Voting, Stacking	Proposed majority voting classifier gives best results as compared with other algorithms.	Majority voting classifier (81.82%)
Mrs. Kelibone Eva Mamabolo ,Dr. Moeketsi Mosia(2020)	Decision Tree(J48), Logistic Regression, Multilayer Perceptron, Naïve Bayes	Class level accuracy should also be considered instead of overall accuracy.	Logistic Regression gives the best accuracy of 57%

V. CONCLUSION

Heart disease is critical disease in this world which has been severely hit by a virus, making heart diseases even riskier and having the knowledge of how machine learning can contribute greatly to predict a disease at an early stage can be used to a great extent here. Some of the best models and automated systems were surveyed and summarized in this paper. Feature selection techniques such as backward selection algorithm, Random Search Algorithm can be used to improve accuracy of the models in so many ways to select the right features. A summary of a few hybrid models is also done. These are the new techniques which are giving better results as compared to classic machine learning algorithms. Hyperparameter tuning is another tool that can greatly help to find optimize values of hyperparameters at which models are giving amazing results.

REFERENCES

[1] T. John Peter; K. Somasundaram "An empirical study on prediction of heart disease using classification data mining techniques" IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012)
[2] Muhammad Fathurachman, Umi Kalsum, Noviyanti Safitri and Chandra Prasetyo Utomo "Heart disease diagnosis using extreme learning based neural networks" 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)
[3] Senthilkumar Mohan, Chandrasegar Thirumalai And Gautam Srivastava "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques " IEEE Access (Volume: 7)
[4] Ashiya Zabeen, Ankur Utsav and Kanhaiya Lal "Detection of Heart Disease Applying Fuzzy Logics and Its Comparison with Neural Networks" 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)
[5] Ming Liu and Younghoon Kim, "Classification of Heart Diseases Based On ECG Signals Using Long Short-Term Memory" 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)
[6] Pahulpreet Singh Kohli, Shriya Arora "Application of Machine Learning in Disease Prediction" 2018 4th International Conference on Computing Communication and Automation (ICCCA)
[7] Rahma Atallah, Amjed Al-Mousa "Heart Disease Prediction Using Machine Learning Majority Voting Ensemble Method" 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)
[8] Amin Ul Haq, Jianping Li, Jalaluddin Khan "Identifying the Predictive Capability of Machine Learning Classifiers for designing Heart Disease Detection System", 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing
[9] Ashir Javeed, Shijie Zhou , Liao Yongjian, Iqbal Qasim , Adeeb Noor, Redhwan Nour "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection" IEEE Access (Volume: 7)
[10] Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms" TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)
[11] Amin Ul Haq and Jianping Li "Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection" Conference: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT)
[12] Sachin Singh, Animesh, and Thomas Penzel "Classification and Detection of Heart Rhythm Irregularities using Machine Learning" 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)
[13] Samir S Yadav, Shivajirao M. Jadhav, Snigdha Nagrale, Niraj Patil "Application of Machine Learning for the Detection of Heart Disease " 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)
[14] Meenu Bhatia and Dr. Dilip Motwani "Use of Ensemblers Learning for Prediction of Heart Disease" 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)
[15] Mrs. Kelibone Eva Mamabolo and Dr. Moeketsi Mosia "Heart Disease Risk Level Prediction: Knitting Machine Learning Classifiers" 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)
[16] https://www.who.int/
[17] https://towardsdatascience.com/
[18] https://www.geeksforgeeks.org/
[19] https://www.kaggle.com/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY