Review of Supervised Machine Learning Algorithms for Heart Disease Prediction

[1]Neetu Kumari, [2]Dr. Anita Ganpati

[1]MTech student, [2]Professor, Computer Science and Engineering, Himachal Pradesh University, Shimla, India. [1]thakurneetu707@gmail.com, [2]anita.ganpati@gmail.com

Abstract **Heart disease is one of the leading causes of life complications and consequences death. Manual heart disease diagnosis and treatment by doctors is a complex and critical task and results in taking plenty of tests. Specifically in third world countries, due to the scarce availability of skilled professionals and diagnostic equipments required for befitted prognostication of the patients. Due to plenty of risk factors conducive to heart disease like high blood pressure and cholesterol, chest pain, uncontrolled diabetes etc it becomes difficult to predict it early. Hence require having a system to predict it as early as possible. For this purpose, various techniques in machine learning have been employed by different researchers. This paper presents the summary of various supervised machine learning algorithms used for predicting the heart disease.**

Keywords — Classification algorithms, Heart disease, machine learning, prediction

I. INTRODUCTION

Heart is a significant part of our body which is responsible for supplying the blood to all over the body. Blood provides our body with essential oxygen and nutrients. It also throws garbage. Fortunately there are best prevention strategies available for reducing the risk of heart disease occurrence such as maintaining healthy diet, getting regular physical activities, keeping blood pressure and cholesterol under control, quit smoking, alcohol and tobacco. If some of them found abnormal then in order to predict the risk of heart disease early, several machine learning techniques are deployed by many researchers are studies in this paper.

Machine learning is the field of study that makes easier to train machines about different circumstances without programmed explicitly and hence machines become capable at predicting after-effects. Machine learning and data mining technology plays a very important role in medical data Analysis and knowledge extraction [1]. The various classification algorithms which were used by different researchers as a solution for predicting the heart disease are discussed in this study.

This paper is organized as follows, Section II introduces the various supervised machine learning algorithms, Section III summarizes recent related work done in this domain and finally the paper is described briefly with the conclusion at the end in Section IV.

II. SUPERVISED MACHINE LEARNING ALGORITHMS

The classification task is used to predict subsequent cases based on past information [2]. If there are only two possible outcomes of a problem then it is known as Binary Classification as Yes or No, if there is a possibility of occurrence of more than two outcomes then that is known as Multi-class Classifier. There are numerous machine learning classification algorithms used by most of the researchers to make accurate diagnoses in heart disease such as Logistic Regression, KNN, Support vector machines, Naïve Bayes, Decision Trees Classification, Random Forest Classification, etc for classifying heart disease. Some important algorithms are discussed here.

A. LOGISTIC REGRESSION

It is a supervised machine learning algorithm which is discriminative and appropriate one when the dependent variable has a binary outcome [3]. This method is mainly used for predicting the dependent binary variable's outcome by using independent features (covariates). This algorithms is preferred when the dependent variable contains only two values, like 0(yes) and 1(no). It is often seen that the target parameter is discrete, taking one or two possible values [4]. so over the last decade Logistic Regression has become the standard method of analysis in this field in many areas [5]. The most common regression methods is Least Squares [6].

 DOI : 10.35291/2454-9150.2022.0142

B. K-Nearest Neighbor (KNN)

KNN is one of the simplest supervised learning techniques. It classifies the objects in the dataset based on the classes of their nearest neighbors. KNN predictions consider that objects near each other are identical [7]. KNN is a costly method to categorize new objects. It is commonly used for its easy of interpretation and low calculation time [8]. To select the value of k in KNN is crucial task. Many researchers modified this algorithm depending upon their needs to solve respective problems. Z. Yunliang, Z. Lijun, Q. Xiaodong and Z. Quan [9] combined KNN algorithm with K-variable algorithm and weighting algorithm, and experienced improvement in the classification of text. Another modified KNN algorithm was proposed by blending classification and clustering into traditional KNN which performed better than traditional one [10]. This algorithm is used in different fields like Pattern Recognition, Cancer Diagnosis, and Text Classification etc. It is a slow learning model since it learns nothing during the training phase but learns only during the testing phase.

C. Support Vector Machine (SVM)

SVM is one of the most popular large - margin classifier which follows structural risk minimization principle [11]. It was developed in statistical learning theory and after that it was used in machine learning, signal processing and statistics [12]. It is employed by many researchers to solve many real life problems. SVM models and Multi Layer Perceptron are somewhat similar [13]. For performing the classification it builds one or more than one hyperplane in the high-dimensional feature space [14]. There are two types of SVM which are linear SVM and non-linear SVM.

D. Naïve Bayes

This classification algorithm is a supervised learning algorithm, and is based on Bayes theorem [15]. Naïve Bayes considers every feature co-relational independent that means changes in one doesn't affect another. It has been used in many complex disease predictions like heart disease, diabetes, liver disease, dengue disease and hepatitis disease [16]. Naïve Bayes classifier requires least training data than most of other classifiers [17]. The formula for Bayes theorem is as

$$P(X/Y) = \frac{P(Y/X) * P(X)}{P(Y)}$$

Where P(X/Y) is the Posterior probability, P(Y/X) is the Likelihood probability, P(X) is the Prior probability, P(Y) is the Marginal probability.

E. Decision Trees

Decision tree is a supervised learning technique that can classify both numerical and categorical features. They perform classification on instances by arranging them based on their feature values [13]. Since it depicts the outcome in a tree-like structure therefore it becomes very easy to understand them [18]. It consists of number of nodes and these nodes can further be categorized as root nodes, internal nodes and leaf nodes. Each internal node points to a test on target variable; each leaf node depicts the target feature's value or represents a class of the dataset. The working mechanism of decision tree for predicting the class includes starting from the root node of the tree and based on the result of comparison of attribute to the root node the procedure of either splitting or jumping is taken and move through until it reached the leaf node. The process of splitting the datasets into various subsets continues until possible values of attributes are reached. Decision tree are effective technique used for the purpose of decision-making.

F. Random Forest

Random Forest is a supervised learning algorithm that is used for classification and regression both. The numbers of trees create a forest in this model. Each tree in this algorithm satisfies the expectation of a class with the most votes is transformed into a model's prediction. In the random forest, averaging the predictions of several weak classifier trees of the forest results in a stronger classifier to give the most accurate prediction results. It is an example of the ensemble method [19]. This model takes less time for training than other models, which makes it an efficient one.

III. Literature Review

Heart disease prediction with the help of supervised machine learning approach is a field of research that has gained a lot of attention in the last decades. In this section the related work done in the domain is presented of heart disease diagnosis by other researchers.

A. Dwivedi [20] presented a framework for recognition of heart disease quickly out of huge sample. The performance of six supervised machine learning algorithms was analyzed by him namely Artificial Neural Network (ANN), Support Vector Machine, Naïve Bayes Classifier, Logistic Regression Classifier, k- Nearest Neighbor and Classification Trees. StatLog heart disease dataset from UCI laboratory was used which contains 13 features and 270 samples. It was examined that logistic regression method performed better with 85% accuracy followed by ANN with 84% accuracy.

S. Hasan, M. Mamun, M. Uddin, and M. Hossain [21] performed comparative analysis of supervised classification approaches for heart disease prediction. In this study less needed features were removed by using info gain feature selection algorithm. The algorithms used include KNN, Guassian Naïve Bayes, Decision Tree (ID3), Logistic Regression and Random Forest. To evaluate the performance of these algorithms, various measurement metrics were used like precision, ROC curve, recall,

 DOI : 10.35291/2454-9150.2022.0142

sensitivity, specificity and F1-score.Maximum accuracy 92.76% was achieved by using Logistic Regression followed by Random Forest with 92.1% accuracy. Except KNN all the algorithms secured 90% above accuracy. They have used Cleveland Heart Disease dataset from UCI machine learning repository which consists of 14 attributes and 303 records.

S. Bashir, Z. Khan, F. Khan, A. Anjum, and K. Bashir [22] focused on improving the heart disease prediction using feature selection algorithms and then trained the models as Decision Tree, Logistic Regression, Naïve Bayes, Random Forest and Logistic Regression,(SVM) by using 5-fold cross validation and after that their accuracy is checked. They obtained dataset from UCI laboratory which consisted of more than 300 attributes but they reduced it to 14 and for data analysis Rapid Miner tool was used. Results of their experiment showed that Logistic Regression (SVM) presented highest accuracy and the better performance. The accuracy of Logistic Regression (SVM) was obtained as 84.85% followed by Naïve Bayes with 84.24% accuracy.

R. Bharti et al. [23] studied the case of heart disease prediction using by using a combination of Machine Learning and Deep Learning to compare the results and analysis of the UCI Machine Learning Public Health Heart Disease dataset. The sample was composed originally 76 attributes but of out of these only 14 relevant attributes were used. The results obtained are validated by suing accuracy and confusion matrix. To handle undesired features Isolation Forest method was used. The classifiers used to do experimentation were Logistic regression, KNN, SVM, Random Forest, Decision tree and Deep Learning (DL).Initially machine learning algorithms are applied and then deep learning is used to check the difference in their performance. In total three approaches were used, in the first one dataset is used directly, in the second one, only feature selection algorithm is used without outlier detection and in the last approach not only the dataset is normalized but outlier detection, removal and feature selection algorithm was also employed. The result obtained in third approach was quite better than other two. The highest accuracy 94.2%was obtained using deep learning approach with 83.1 Specificity and 82.3 Sensitivity. In this study it is also explained that how it can be combined with some multimedia technology e.g. mobile devices.

V. Sharma and S. Yadav [2] proposed a heart disease prediction using machine learning algorithms such as Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision tree. They used a dataset from UCI that contains 14 different features with 1025 instances. At the end they found that Random Forest better acquired better prediction results with 99% accuracy followed by SVM which gave 98% accuracy and Decision Tree performed worst with 85% prediction accuracy.

D. Shah, S. Patel, and S. Kumar [24] compared the performances of four supervised machine learning algorithms such as Naïve Bayes, decision tree, K-nearest neighbor, and Random Forest models. The Cleveland dataset of UCI repository of heart disease patients which commonly consist 76 attributes and 303 instances but only 14 attributes were used for training and testing all these models. The experiment was conducted on WEKA tool and evaluation was noted using Python programming based on accuracy score and it was found that KNN appeared the best with 90.78% accuracy followed by Naïve Bayes algorithm with 88.15% accuracy.

A. Otoom, E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour [25] proposed an effective a real-time diagnosis and monitoring system for heart disease. It was composed of two components named as diagnosis component and monitoring component. The first component i.e. diagnosis component of the system is responsible for diagnosis of heart disease whereas another component was an inexpensive wearable sensor that monitors the frequency of heart and wirelessly send the recorded signal to a mobile device. They compared three supervised machine learning techniques to find the efficient one for heart disease diagnosis named as BayesNet, Support Vector Machine (SVM) and Functional Trees with and without feature selection. For model training and testing Cleveland heart disease data set from the UCI Machine Learning Repository was used that consisted of 303 instances with 76 features; however, only 13 relevant features were used. They conducted two experiments for Diagnosis and Monitoring components on Nokia Lumia 520 mobile phone. The results show that BayesNet and Function Tree were found to be the algorithms with 84.5% highest accuracy.

S. Arunachalam [26] proposed cardiovascular disease prediction model using Machine Learning algorithms. For conducting this study the Cleveland, Hungarian, Switzerland, Long Beach VA heart disease database present in the UCI machine Learning Repository was used which consisted of total 300 instances and 76 attributes but only 14 attributes were considered in this experiment. 80% data was used for training the models and 20% data for testing them from the dataset. Several classification algorithms such as: Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, Extremely Randomized Trees Classifier (Extra Trees Classifier), Logistic Regression and Multi-Layer Perceptron (MLP) Classifier were trained and tested on the dataset further they are analyzed against accuracy, sensitivity and specificity in order to obtain the performance of the diagnosis model. From the experimental results it can be concluded that SVM and MLP provided the highest accuracy of 91.7%. To display the outcome of entered values based on available parameters of the dataset, a front-end system was also proposed.

 DOI : 10.35291/2454-9150.2022.0142

C. Latha and C. Jeeva [15] investigated a method known as ensemble classification for improving the performance of weak models by merging numerous classifiers to make more efficient the heart disease prediction. They utilized The Cleveland heart dataset from the UCI machine learning repository for conducting the experiment**.** The dataset contained 303 instances with 14 attributes. Various classification algorithms such as: Bayes Net, Naïve Bayes, Random Forest, C4.5, MLP and PART (Projective Adaptive Resonance Theory) were trained and tested through the dataset and then they were improved by applying bagging, boosting, stacking, voting and further they were enhanced by feature selection. At the end the results showed that majority vote with NB, BN, RF and MP achieved 85.48% accuracy.

M. Karthikeyan, C. Myakala, and S. Chappidi [27] had proposed a user interface to enter the input values based on used dataset and displaying the results accordingly. As XGBoost (Extreme Gradient Boosting) Classification algorithm showed the highest accuracy therefore this algorithm is used on the user interface for showing the outcomes against heart disease prediction with at most accuracy. Dataset was collected from the University of California, Irvine's machine learning repository for conducting this experiment. The dataset had 303 samples and 14 attributes out of which 11 attributes detection of heart disease whereas 2 attributes were used for knowing the data of patient. The classification algorithms that were analyzed in this experiment are: Logistic regression, Naïve Bayes, XGBoost and Decision Tree Gini Index. The highest accuracy 90.46% was obtained by using XGBoost algorithm.

M. Ali et al. [28] studied various supervised machine learning classifiers such as KNN, Random Forest, Decision Tree, AdaboostM1(ABM1), Logistic Regression and MLP against the prediction of heart disease. The dataset was taken from Kaggle. For experimentation, WEKA (3.8.3) tool was utilized and for Exploratory Data Analysis(EDA) and visualization, pyton(3.8.5) was used. Except k-Nearest Neighbor and Multi Layer Perceptron, feature importance scores were estimated because these two algorithms don't generate or support feature importance score. The ranking of features was done in order to identify the most important features to estimate the heart disease prediction. KNN, Decision Tree and Random Forest provided 100% accuracy, based on the dataset used.

E. Hashi and M. Zaman [29] developed a Hyperparameter Tuning based machine learning approach of heart disease prediction. They implemented several classification models such as KNN, Logistic regression, SVM, Decision tree, and Random Forest classifier. For tuning the hyperparameters on these classifiers the grid search approach was utilized. The researchers had used the Cleveland heart disease dataset for training and testing the classification algorithms which was consisted of 75 parameters (only 14 were used) and 303 instances. The main objective behind using Hyperparameter tuning approach was to increase the accuracy of the models mentioned before. At the end the performance of these models with and without hyperparameter was evaluated and it was examined that accuracy of LR, KNN, SVM, DT, and RF classifiers increased from 88.52%, 90.16%, 88.52%, 81.97%, and 85.25 to 90.16%, 91.80%, 90.16%, 86.89%, and 85.25% respectively.

P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha [30] proposed a machine learning framework to predict heart disease occurrence chances. Five machine learning algorithms such as Random Forest, Naïve Bayes, Support Vector Machine, Hoeffding Decision Tree, and Logistic Model Tree (LMT) were trained and tested through the Cleveland dataset. The performance of all these algorithms was examined and then individually after 3applying bagging and boosting algorithms to check the increased accuracy. Finally, the feature selection algorithm was used and only the selected features were used to train and test these models, and after that, the visualizations are drawn. In the end, it was discovered that the Random Forest algorithm outperformed with 95.08% accuracy other used algorithms.

G. Kumar, D. Kumar, K. Arumugaraj and V. Mareeswari [31] used different methods and compared their accuracies in order to develop a prediction model for cardiovascular disease by using the most accurate method. They also designed a user interface to provide convenience to the users. The methods used by them in this study were Random Forest, SVM, Naïve Bayes, Logistic Regression, and GBM. They used a dataset from the UCI Machine learning repository of heart disease in which there were 920 instances with 76 attributes out of which 14 attributes were used. R language was used for implementation and graphical visualization of results obtained. From the results obtained it was found that Logistic Regression outperformed the other four models with 91.61% accuracy followed by the Naïve Bayes model with 90.95% accuracy.

Y. Khourdifi and M. Bahaj [32] proposed a hybrid method for not only the prediction of heart disease but classification too. In this study, they used five classification algorithms as KNN, SVM, Naïve Bayes, Random Forest Classifier, and MLP. Further optimization was performed by combining two approaches as Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). The dataset from the Cleveland database (UCI) was used in this study with 14 attributes. The algorithms were optimized by using a feature selection method known as Fast Correlation Based Filter (FCBF), the combination of PSO and ACO. The implementation was done using the

 DOI : 10.35291/2454-9150.2022.0142

WEKA tool. The results obtained proved that the performance of proposed classifiers outperformed many recent classifiers. The highest accuracy was given by KNN that was 99.65% followed by MLP with 99.60% accuracy.

F. Alotaibi [33] implemented and compared five machine learning models namely Decision Tree(DT), Logistic Regression (LR), Random Forest(RF), Naïve Bayes(NB), and Support Vector Machine (SVM) for improving the heart failure prediction accuracy. The dataset used in this study was taken from the Kaggle (Cleveland data) and out of which fourteen attributes were used along with 303 instances. To enhance the accuracy of the machine learning models for heart failure detection several steps they had followed. Initially, data preprocessing was performed on a dataset in order to detect and remove null values, noisy data, etc. As the sample was of the dataset was very small, so to avoid biases the random number generation approach was used for each feature. Further for the data cleaning, they used the Rapid Miner tool. Finally, the models are trained using 10-fold cross-validation and are implemented using Rapid Miner. The results depicted that the Rapid Miner tool helped to enhance the heart failure prediction accuracy which was lower when analyzed using the WEKA tool and MATLAB. Decision Tree achieved the highest accuracy of 93.19% followed by SVM with 92.3% accuracy.

S. Patel, T. Upadhyay and S. Patel [34] compared three algorithms namely J48, logistic model tree and random forest algorithm for finding the best one in order to build the heart disease prediction model. They selected the Cleveland dataset from UCI machine leaning repository of heart disease. There were 303 instances and 76 parameters but only 13 parameters were used in this study. The algorithms were implemented using WEKA 3.6.10 tool. They found that J48 algorithm performed better with lesser time in building. Further they had evaluated the results for all the models with and without error pruning. It was observed that J48 achieved highest accuracy that was 56.76% with error pruning and it took 0.04 seconds to build.

K. Amen, M. Zohdy and M. Mahmoud [35] classified heart disease into five stages namely no heart disease, first stage, second stage, third stage and severe heart disease. The Cleveland dataset of heart disease from UCI machine learning repository was employed which consisted of fourteen attributes and 303 samples. Further they used five machine learning classifiers such as Logistic Regression, Support Vector Machine, Random Forest, Gradient Tree Boosting and Extra Random Forest with hyper parameters. These classifiers were implemented using Python language and evaluated by using some parameters like precision, recall, accuracy and F measure. The maximum accuracy of 82% was shown by Logistic Regression.

Y. Khourdifi and M. Bahaj [36] proposed optimized KNN model for heart disease classification. Two optimization approaches named as Particle Swarm optimization (PSO) and Ant Colony Optimization (ACO) were used. They used the heart disease dataset from Cleveland Clinic Foundation of UCI repository. The Fast Correlation-Based Feature Selection (FCBF) method was used for most relevant feature extraction. Further they proposed a modified KNN by using aforementioned optimization techniques and the resulting model was named as PA-KNN. Six supervised classification algorithms were built using WEKA in this study like KNN, SVM, RF, NB, MLP and PA-KNN. By using 10-fold cross validation, these models were trained and tested. On comparison it was discovered that the proposed algorithm named as PA-KNN outperformed other. The highest accuracy of 99.7% was achieved by PA-KNN.

V. Jayaraman and H. Sultana [37] introduced a modified algorithms for reducing the features of the dataset for acquiring more accuracy in heart disease prediction. The proposed algorithm was formed by combining two different algorithms such as gravitational cuckoo search and particle bee optimized associative memory neural network. The dataset was used from heart disease UCI repository of machine learning which consisted of 76 features and 303 instances. Machine learning algorithms implemented and compared in this study were Bat- based back propagation (BAT-BP) algorithm, Genetic algorithm optimization of a convolutional neural network (GA-CNN), Ant colony optimization neural networks (ACONN) and Particle bee optimized associative memory neural network (PBAMNN). The maximum accuracy of 99.85% was shown by proposed algorithm PBAMNN followed by ACONN with 98.61%.

A. Javeed et al. [38] introduced an intelligent diagnostic system for heart disease. For feature selection, the proposed system had used RSA (Random Search Algorithm) and for prediction of heart failure, Optimized Random Forest Model was used. The optimization of introduced diagnostic system was done by using Grid Search Algorithm. The experimentation was done in two stages. Initially only RF model was built and then in the second stage, the development of proposed Random Search Algorithm based RF model was done. For conducting this study the Cleveland dataset was utilized. In total, six machine learning algorithms were compared namely adaboost ensemble model, extra tree ensemble model, random forest, SVM model, SVM (RBF) model and proposed (RSA-RF) model. The proposed RSA-RF model achieved highest accuracy of 93.33%.

S. Mohan, C. Thirumalai, and G. Srivastava [39] proposed a model named HRFLM (Hybrid Random Forest with Linear Model) for heart disease prediction by using hybrid machine learning algorithms. The new model is

 DOI : 10.35291/2454-9150.2022.0142

produced by combining the features of Random Forest (RF) and Linear Method (LM). Artificial Neural Network (ANN) with back propagation was used in HRFLM to fed the data. The accuracy exhibited by this model was 88.7%.

Table 1 summarizes various supervised machine learning classification algorithms described above for predicting the heart disease. It also showing the algorithm with the highest accuracies, dataset used and environment employed by the researchers to conduct the experiment. The table contains the following elements.

Year of publication: This contains the year information of the research paper publication.

Authors: This shows the authors of the research paper.

References: This field depicts the reference number to the corresponding research paper.

Technique: This column is representing the type/types of classification algorithms used for the experimentation.

Maximum accuracy achieved: This shows the algorithm name with the highest accuracy and the percentage of accuracy achieved in the experiment.

Dataset: This column is containing the information about the dataset name and the directory from where it is taken.

Environment: This field is giving the details about the framework, tool and the programming language used for conducting the studies.

Table 1: Summary of supervised machine learning algorithms

Year of publication	Authors	References	Technique	Maximum Accuracy Achieved	Dataset	Environment
2015	A. Otoom et al.	**[25]**	BayesNet, SVM and FT	BayesNet and FT(84.5%)	Cleveland (UCI)	Nokia lumia 520 mobile phone(For diagnosis) , Pulse Sensor AMPED (For monitoring heart rate)
2016	A. Dwivedi	**[20]**	SVM, ANN, NB, LR, KNN	LR(85%)	StatLog (UCI)	Weka (3.6)
2016	S. Patel	**[34]**	J48, LMT, RF	J48(56.76%)	Cleveland(UCI)	Weka (3.6.10)
2018	S. Hasan et al.	**[21]**	KNN, DT, Guassian NB, LR, RF	LR (92.76%)	Cleveland (UCI)	Anaconda Python(Spyder 3.6)
Year of publication	**Authors**	**References**	**Technique**	**Maximum Accuracy Achieved**	**Dataset**	**Environment**
2018	D.G. et al.	**[31]**	LR, RF, NB, GBM, SVM	LR (91.61%)	UCI	R language for statistical computation
2019	S. Mohan et al.	**[39]**	DT, LM, SVM, RF, NB, NN, KNN, HRFLM (proposed)	HRFLM (88.7%)	Cleveland (UCI)	R studio rattle
2019	Javeed et al.	**[38]**	adaboost, ET, RF, SVM, SVM (RBF) and proposed (RSA-RF)	RSA-RF (93.33%)	Cleveland(UCI)	Python programming language
2019	V. Jayaram-an	**[37]**	BAT-BP, GA-CNN, ACONN and PBAMNN	PBAMNN (99.85%)	Bouckaert and Frank (UCI)	MATLAB
2019	Y. Khourdifi and M. Bahaj	**[32]**	KNN, SVM, RF, NB and MLP	KNN (99.65%)	Cleveland (UCI)	Weka
2019	F. Alotaibi	**[33]**	DT, LR, RF, NB, SVM	DT (93.19%)	Cleveland(UCI)	Rapid Miner
2019	C. Latha and S. Jeeva	**[15]**	BayesNet, NB, RF, C4.5, MLP and PART	Majority vote with NB, BN, RF and MLP (85.48%)	Cleveland (UCI)	Weka(For classification)
2019	Y. Khourdi and M. Bahaj	**[36]**	KNN, SVM, RF, NB MLP, PA-KNN	PA-KNN (99.7%)	Cleveland (UCI)	Weka
2020	K. Amen et al.	**[35]**	SVM, RF, LR, GTB and ERF	LR (82%)	Cleveland (UCI)	Python programming language
2020	V. Sharma et al	**[2]**	RF,DT, NB, SVM	RF (99%)	Cleveland (UCI)	Weka
2020	D. Shah et al.	**[24]**	NB, KNN, DT, RF	KNN(90.7)	UCI	Weka and python
2020	**S.** Arunacha-lam	**[26]**	GB, RF, SVM, ET,LR ,MLP	SVM and MLP (91.7%)	Cleveland, Hungarian, Switzerland, Long Beach VA (UCI)	Python programming language and front end systems was also deployed

 DOI : 10.35291/2454-9150.2022.0142

2020	M. Karthike-yan et al.	**[27]**	LR, NB, XGBoost and DT Gini Index	XGBoost (90.46%)	University of California, (Irvine's)	Flask(For building web application)
2020	E. Hashi and Md. Zaman	**[29]**	LR, KNN, SVM, DT, and RF	KNN (91.80%)	Cleveland(UCI)	Python programming language
2020	P. Motarwar et al.	**[30]**	RF, NB, SVM, HDT and LMT	RF (95.08%)	Cleveland(UCI)	Python programming language
2021	R. Bharti et al.	**[23]**	LR,KNN,SVM,RF,DT,DL	DL (94.2%)	Cleveland, Hungary, Switzerland, and Long Beach V(UCI)	Python programming language
2021	Md. Ali et al.	**[28]**	KNN, RF, DT, ABM1, LR and MLP	KNN, DT and RF (100%)	Kaggle (UCI)	WEKA 3.8.3 (For EDA) and Pyton 3.8.5(For visualization)

I. Results And Analysis

The literature was reviewed on the heart disease prediction from the year 2015 to 2021. Almost every researcher experimented with the Cleveland dataset from UCI directory of heart disease prediction. For the prediction of the occurrence of heart disease several algorithms performed well. It was observed that mostly the Random Forest, K-Nearest Neighbor and Logistic Regression offered better accuracy as compared to others. Naïve Bayes algorithm was used widely for heart disease prediction, but it failed to provide expected results due to its strong feature independence assumptions. SVM classifier was utilized in almost every study along with other classifiers, but it is not suitable for large data sample. Decision tree classifier was also used and offered good results but due to over-fitting issue it is not widely used. Deep Learning, one of the machine learning techniques is also very useful for prediction. In order to use Deep Learning method the data sample must have large instances of data to provide better performance. Deep Learning requires comparatively more expensive processing and storage resources.

II. Conclusion

From the vast literature review, it was found that the different accuracy score of classifiers in predicting heart disease was obtained in each study. It was observed that the algorithm performance varied with the dataset, the number of features and instances of the dataset, the environment, preprocessing of the dataset, classification algorithms, the training of algorithms, and the feature selection algorithm used. We found that mostly the Cleveland dataset from the UCI repository was used with 303 instances and 14 features. For implementation WEKA tool was utilized mostly. Three classification algorithms named Random Forest, K-Nearest Neighbor, and Logistic Regression provided better accuracy in most of the studies.

In order to build an effective and intelligent heart disease prediction system, heart disease samples from different countries should be collected. The proper preprocessing of the dataset is a must to avoid over-fitting and under-fitting issues. By training the different classification algorithms with the vast geographical heart disease samples, the results will be more accurate and practical. We conclude that algorithms should be trained and tested with more instances and lesser features are a good practice.

Acknowledgment

Firstly I'm thankful to the God, for bestowing the wisdom, strength and good health to me in order to accomplish this task.

Secondly, I would like to thank my passionate and supportive supervisor who guided me in every possible way. Her guidance made this work possible. Her valuable suggestions helped me in each stage.

References

[1] L. Yahaya, N. David Oye, and E. Joshua Garba, "A comprehensive review on heart disease prediction using data mining and machine learning techniques," *Am. J. Artif. Intell.*, vol. 4, no. 1, p. 20, 2020, doi: 10.11648/j.ajai.20200401.12.

[2] V. Sharma, "Heart disease prediction using machine learning techniques," *2020 2nd Int. Conf. Adv. Comput. Commun. Control Netw.*, pp. 177–181, 2020.

[3] K. Raza, "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule", Elsevier Inc., 2019.

[4] L. Gladence, M. Karthi, and V. Anu, "A statistical comparison of logistic regression and different bayes classification methods for machine learning," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 14, pp. 5947–5953, 2015.

 DOI : 10.35291/2454-9150.2022.0142

[5] A. J. Scott, D. W. Hosmer, and S. Lemeshow, "Applied logistic regression.," *Biometrics*, vol. 47, no. 4, p. 1632, 1991, doi: 10.2307/2532419.

[6] F. Khan, I. Khan, N. Taj, A. Raghavendra, and D.Chethan, "Heart disease prediction using logistic regression algorithm," *12th Int. Conf. Adv. Comput. Control. Telecommun. Technol. ACT 2021*, vol. 2021-Augus, no. 2, pp. 391–395, 2021.

[7] K. Taunk, "A brief review of nearest neighbor algorithm for learning and classification," *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. Iciccs, pp. 1255–1260, 2019.

[8] A. Moldagulova and R. Sulaimen, "Using KNN Algorithm for Classification of Textual Documents," *017 8th Int. Conf. Inf. Technol. Using*, pp. 665–671, 2017.

[9] Y. Zhang, L. Zhu, X. Qiao, and Q. Zhang, "Flexible KNN algorithm for text categorization by authorship based on features of lingual conceptual expression," *2009 WRI World Congr. Comput. Sci. Inf. Eng. CSIE 2009*, vol. 2, pp. 601–605, 2009, doi: 10.1109/CSIE.2009.363.

[10] S. Taneja, C. Gupta, K. Goyal, and D. Gureja, "An enhanced K-nearest neighbor algorithm using information gain and clustering," *Int. Conf. Adv. Comput. Commun. Technol. ACCT*, no. February 2016, pp. 325–329, 2014, doi: 10.1109/ACCT.2014.22.

[11] Y. Zhang, "Support vector machine classification algorithm and its application," *Commun. Comput. Inf. Sci.*, vol. 308 CCIS, no. PART 2, pp. 179–186, 2012, doi: 10.1007/978-3-642-34041-3_27.

[12] Y. Ma and G. Guo, *Support vector machines applications*, vol. 9783319023. 2014.

[13] F. Osisanwo, "Supervised machine learning algorithms: classification and comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.

[14] S. Ghosh, S. Mondal, and B. Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier," *1st Int. Conf. Autom. Control. Energy Syst. - 2014, ACES 2014*, pp. 1–4, 2014, doi: 10.1109/ACES.2014.6808002.

[15] C. Latha and S. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. June, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[16] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 09, no. 01, pp. 1–16, 2017, doi: 10.4236/jilsa.2017.91001.

[17] P. Kaviani and S. Dhotre, "Short survey on naive bayes algorithm," *Int. J. Adv. Eng. Res. Dev.*, vol. 4, no. 11, pp. 607–611, 2017. [18] B. Deshpande, "Decision tree digest - understand, build and use decision trees for common business problems with RapidMiner," *SimaFore*, 2014.

[19] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[20] A. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, 2016, doi: 10.1007/s00521-016-2604-1.

[21] S. Hasan, M. Mamun, M. Uddin, and M. Hossain, "Comparative analysis of classification approaches for heart disease prediction," *Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018*, 2018, doi: 10.1109/IC4ME2.2018.8465594.

[22] S. Bashir, Z. Khan, F. Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," *2019 16th Int. Bhurban Conf. Appl. Sci. Technol.*, pp. 619–623, 2019, doi: 10.1109/IBCAST.2019.8667106.

[23] R. Bharti et al., "Prediction of heart disease using a combination of machine learning and deep learning," vol. 2021, 2021.

[24] D. Shah, S. Patel, and S. Kumar, "Heart disease prediction using machine learning techniques," *SN Comput. Sci.*, vol. 1, no. 6, pp. 1–6, 2020, doi: 10.1007/s42979-020-00365-y.

[25] A. Otoom, E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease," *Int. J. Softw. Eng. its Appl.*, vol. 9, no. 1, pp. 143–156, 2015, doi: 10.14257/ijseia.2015.9.1.12.

[26] S. Arunachalam, "Cardiovascular disease prediction model using machine learning algorithms," no. June, 2020.

[27] M. Karthikeyan, C. Myakala, and S. Chappidi, "Heart attack prediction using XGBoost," vol. 29, no. 6, pp. 2392–2399, 2020.

[28] M. Ali et al., "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, no. May, p. 104672, 2021, doi: 10.1016/j.compbiomed.2021.104672.

 DOI : 10.35291/2454-9150.2022.0142

[29] E. Hashi and M. Zaman, "Developing a hyperparameter tuning based machine learning approach of heart disease prediction," *J. Appl. Sci. Process Eng.*, vol. 7, no. 2, pp. 631–647, 2020, doi: 10.33736/jaspe.2639.2020.

[30] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive approach for heart disease prediction using machine learning," *Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020*, 2020, doi: 10.1109/ic-ETITE47903.2020.242.

[31] G. Kumar, D. Kumar, K. Arumugaraj and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," *2018 Int. Conf. Curr. Trends Towar. Converging Technol.*, pp. 1–7, 2018.

[32] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 242–252, 2019, doi: 10.22266/ijies2019.0228.24.

[33] F. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 261–268, 2019, doi: 10.14569/ijacsa.2019.0100637.

[34] S. Patel, T. Upadhyay and S. Patel, "Heart disease prediction using machine learning and data mining technique," no. March, 2016, doi: 10.090592/IJCSC.2016.018.

[35] K. Amen, M. Zohdy, and M. Mahmoud, "Machine learning for multiple stage heart disease prediction," pp. 205–223, 2020, doi: 10.5121/csit.2020.101118.

[36] Y. Khourdi and M. Bahaj, "K-nearest neighbour model optimized by particle swarm optimization and ant colony optimization for heart disease classification," vol. 53, pp. 215–224, 2019, doi: 10.1007/978-3-030-12048-1.

[37] V. Jayaraman and H. Sultana, "Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification," *J. Ambient Intell. Humaniz. Comput.*, vol. 0, no. 0, p. 0, 2019, doi: 10.1007/s12652-019-01193-6.

[38] A. Javeed et al., "An intelligent learning system based on random search algorithm and optimized random forest Model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

[39] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using yybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

 DOI : 10.35291/2454-9150.2022.0142