International Journal of Electrical and Electronics Engineers
Volume 16, Issue No. 01, Jan-June 2024
ISSN (O) 2321-2055
ISSN (P) 2321 -2045

MACHINE LEARNING MODELS TO PREDICT HEART DISEASE

Guani A Sabi[1], Dr.G.Maria Kalavathy, M.E, Mba, Ph.D[2]

[1]Department of Computer Science andEngineering Omr Chennai-600119

guanisabi7@gmail.com

[2]Department of Computer Science andEngineeringOmr Chennai-600119

mariakalavathy.yahoo.in

Abstract—

This paper focuses on predicting heart disease using various models. By utilizing the patient's medical records, a new system is proposed to predict the chances of a person contracting heart disease. Attributes such as age, blood pressure, pulse rate, diabetes, etc.…, which is used to predict the risk of heart disease in a person. The main focus of the system is to make use of data analysis to predict the presence of the disease and the level of disease among patients. This paper compares the accuracy of machine learning algorithms for predicting Heart disease, for these algorithms are k-nearest neighbor, decision tree, linear regression, and support vector machine(SVM) by using the UCI repository dataset for training and testing.

Keywords— Blood pressure, pulse rate, heartdisease, SVM, KNN, Decision tree, Linear Regression.

I. Introduction

Predicting heart disease (CVD) using machine learning models is a critical endeavor in healthcare, aiming to identify individuals at risk of CVD before it manifests, thereby allowing for early intervention and prevention. The introduction to this field emphasizes the importance of early detection and prevention of CVD, which is a leading cause of death worldwide. Machine learning offers a promising approach to this challenge by analyzing complex datasets containing patient information, such as demographics, medical history, lifestyle factors, and genetic data, to predict the likelihood of developing CVD. Machine learningmodels can process vast amounts of data to identify patterns and correlations that may not be immediately apparent to humans. These models can be trained on datasets that include features such as age, sex, blood pressure, cholesterol levels, smoking status, physical activity, and family history of CVD. The models then use this information to predict whether an individual isat risk of developing CVD, based on their current healthstatus and lifestyle.

The application of machine learning in CVD prediction has several advantages. Firstly, it can handle large datasets efficiently, allowing for a comprehensive analysis of potential risk factors. Secondly, it can update its predictions as new data becomes available, making it a dynamic tool for ongoing assessment. Thirdly, it can

identify subtle patterns that may not be obvious to clinicians, potentially leading to more accurate predictions. However, the use of machine learning models in CVD prediction also presents challenges. These include the need for high-quality, representative datasets, the risk of overfitting or underfitting, and the ethical considerations of usingsensitive health information. Addressing these challenges requires ongoing research and development, as well as collaboration between healthcare professionals, data scientists, and ethicists.

In conclusion, machine learning models offer a powerful tool for predicting heart disease, with thepotential to improve early detection and prevention efforts. By analyzing a wide range of patient data, these models can provide valuable insights that may lead to more effective interventions and personalized care plans for individuals at risk of CVD.

II. LITERATURE SURVEY

1. A. H. M. S. U. Maria Sultana et al Heartdisease is considered one of the major causes of death throughout the world. It cannot be easily predicted bymedical practitioners as it is a difficult task that demands expertise and higher knowledge forprediction. This paper addresses the issue of prediction of heart disease according to input attributes based on data mining techniques. We have investigated heart disease prediction using KStar, J48, SMO, Bayes Net, and Multilayer Perceptron through Weka software. The performance of these data mining techniques is measured by combining the results of predictive accuracy, ROC curve, and AUC value using a standard data set as well as a collected data set. Based on performance factor SMO and Bayes Nettechniques show optimum performances than the performances of KStar, Multilayer Perceptron, and J48 techniques.

2. Yeshvendra Singh et al The scope of Machine Learning algorithms is increasing in predictingvarious diseases. The nature of machine learning algorithms to think like a human being makes this concept so important and versatile. Here the challenge of increasing the accuracy of Heart disease prediction is taken upon. The non-linear tendency of the Cleveland heart disease dataset was exploited for applying Random Forest to get an accuracy of 85.81%. The method of predicting heart diseases using a Random Forest with well-set attributes fetches us more accuracy. Random Forest was built by training 303 instances of data and authentication of accuracy was done using 10-fold cross-validation. By the proposed algorithm for heart disease prediction, many lives could be saved in the future.

3. M. Akhil Jabbar et al Data mining techniques have been widely used to mine knowledgeable information from medical databases. In data mining classification is a supervised learning that can be used to design models describing important data classes, where class attribute is involved in the construction of the classifier. Nearest neighbor (KNN) is a very simple, popular, highly efficient, and effective algorithm for pattern recognition. KNN is a straightforwardclassifier, where samples are classified based on the class of their nearest neighbor. Medical databases are high-volume in nature. If the data set contains redundant and irrelevant attributes, classification may produce less accurate results. Heart disease is the leading cause of death in INDIA. In Andhra Pradesh heart disease was the leading cause of mortality accounting for 32%of all deaths, a rate as high as Canada (35%) and the USA. Hence there is a need to define a decision support system that helps clinicians decide to take precautionary steps. In this paper, we propose a new algorithm that combines KNN with a genetic algorithm for effective classification. Genetic algorithms perform global searches in complex largeand multimodal

landscapes and provide optimal solutions. Experimental results show that our algorithmenhances the accuracy in the diagnosis of heart disease.

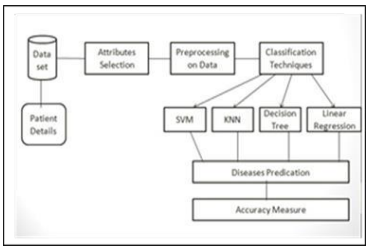
4. Jaymin Patel et al Heart disease has been the main reason for death in the world over the last decade. Almost one person dies of Heart disease aboutevery minute in the United States alone. Researchers have been using several data mining techniques to helphealthcare professionals in the diagnosis of heart disease. However, using data mining techniques can reduce the number of tests that are required. To reduce the number of deaths from heart diseases there has to be a quick and efficient detection technique. Decision Tree is one of the effective data mining methods used. This research compares different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using WEKA. The algorithms that are tested a r e t he J48 algorithm, the Logistic model tree algorithm, and the Random Forest algorithm. The existing datasets ofheart disease patients from the Cleveland database of the UCI repository are used to test and justify the performance of decision tree algorithms. This dataset consists of 303 instances and 76 attributes. Subsequently, the classification algorithm that has optimal potential will be suggested for use in sizeable data. The goal of this study is to extract hidden patterns by applying data mining techniques, that are noteworthy to heart diseases, and to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence.

5. Nikhil Kumar Mutyala et al Data mining is themost popular knowledge extraction method for knowledge discovery (KDD). Machine learning is used to enable a program to analyze data, understand correlations, and make use of insights to solveproblems and/or enrich data for prediction. Data mining techniques and machine learning algorithms play a very important role in the medical area. The healthcare industry contains a huge amount of data. But most of it is not effectively used. Heart disease is one of the main reasons for the death of people in the world. Nearly 47% of all deaths are caused by heart disease. We use 8 algorithms including aDecision Tree, J48 algorithm, Logistic model tree algorithm, Random Forest algorithm, Naïve Bayes, KNN, Support Vector Machine, and Nearest Neighbour to predict heart diseases. The accuracy of the prediction level is high when using a greater number of attributes. We aim to perform predictive analysis using these data mining, and machine learning algorithms on heart diseases analyze the various mining and machine learning algorithms used, and conclude which techniques are effective and efficient.

III. PROPOSED SYSTEM

The proposed approach was applied to the patient details from the dataset and the different attributes are selected from the dataset. The selected attributes are pre-processed. The pre-processed dataset is collected and cleaned, and then different machine learning algorithm models selection in which Linear Regression was used. For focusing on the neighbor selection technique K-Nearest Neighbors Classifier was used, and then a tree-based technique like the Decision Tree Classifier was used. Also, for checking the high dimensionality of the data and handling it, a Support Vector Machine was used. The decision Tree method combination is the XG Boost classifier.

International Journal of Electrical and Electronics Engineers
Volume 16, Issue No. 01, Jan-June 2024
ISSN (O) 2321-2055
ISSN (P) 2321 -2045



Proposed methodology architecture diagram

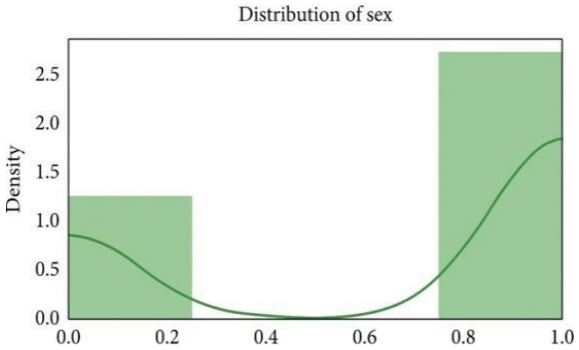
IV. METHODOLOGY DATASET PREPARATION:

The data is collected from Kaggle, and it contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 features. Thedataset is a combination of 4 different databases, but only the UCI Cleveland dataset was used. The target field refers to the presence of heart disease in the patient. The dataset is categorized into two directories training and testing, and it contains subfolders for each image category heart disease and normal. The training and testing dataset consists of a range of each image of heart disease and normal. The diseased images that are predicted as true are separated from the model by checking the model prediction that can either be '0' and be '1' and are provided to the segmentation code where the use of contour marking, and canny edge detection.

PRE-PROCESSING DATA:

The pre-processing data is used to clean the attributes. The next step is the feature selection process and directly applying the data to the machine learning algorithms and the results that were achieved. By using the normal distribution of the dataset to overcome the problem and then applying isolation forest for the outlier detection. Various plotting techniques were used to check the skewness of the data, outlierdetection, and the distribution of the data. All these preprocessing techniques play an important role whenpassing the data for classification or predictionpurposes.

Feature Selection: A critical part of the methodology is the selection of relevant features that contribute significantly to the prediction of CVD. Techniques such as Relief and LASSO feature selection are employed to identify the most informative features, reducing the dimensionality of the dataset and improving the model's performance.



International Journal of Electrical and Electronics Engineers
Volume 16, Issue No. 01, Jan-June 2024
ISSN (O) 2321-2055
ISSN (P) 2321 -2045

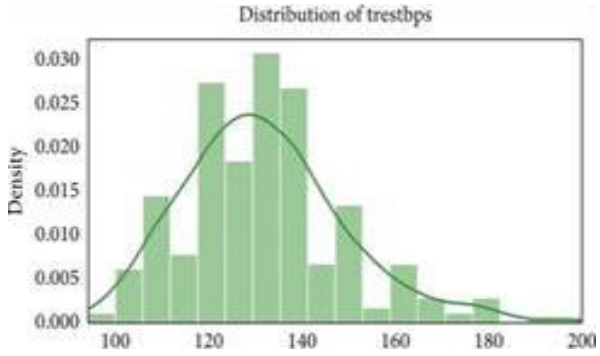


Fig 1 Feature Extraction Result

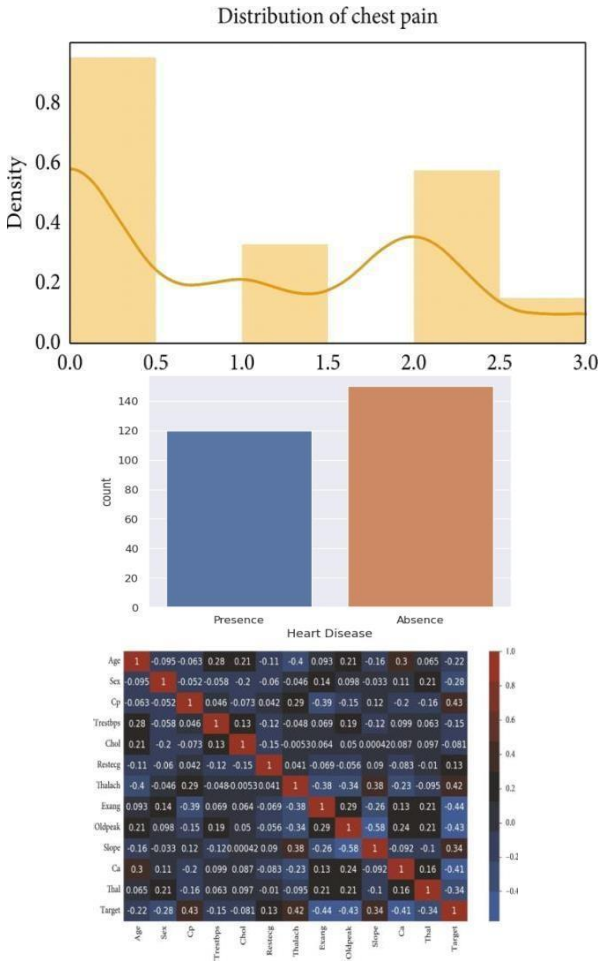


Fig 2 Target class view

Model Training and Evaluation: The selected features and algorithms are used to train the models on the

prepared dataset. The data is typically split into training and testing sets, with a common split ratio being 70% for training and 30% for testing. The performance of the models is evaluated using metrics such as accuracy, sensitivity, and specificity to ensure they are capable of accurately predicting CVD.

Hyperparameter Optimization: To further improve the models' performance, hyperparameter optimization techniques such as grid search and Bayesian optimization are used. These techniques helpin finding the optimal values for the model parameters,which can significantly affect the model's accuracy and generalization ability.

Cross-Validation: To assess the model's generalization ability and to ensure that it performs well on unseen data, k-fold cross-validation is employed. This method divides the dataset into k subsets and trains the model on a k-1 subset while testing it on the remaining subset. This process isrepeated k times, providing a more reliable estimate ofthe model's performance.

Model Fusion: For enhancing prediction accuracy, a stacking model approach is used, which involvescombining the predictions of multiple models (base learners) to make a final prediction. This approach leverages the strengths of each model and can lead to improved performance, especially when the base learners are diverse and accurate.

By following this methodology, machine learning models can be effectively developed and trained to predict CVD with high accuracy. These models can be a valuable tool in healthcare, aiding in early detection and prevention of CVD, thereby potentially savinglives and improving health outcomes.

V. CONCLUSION AND FUTURE ENHANCEMENT

The conclusion of the research on predicting heart disease (CVD) using machine learning models underscores the significant potential of these models inhealthcare, particularly in the early detection and diagnosis of CVD. The study highlights the effectiveness of various machine learning algorithms, including Support Vector Machines (SVM), Gradient Boosting Machines, and Random Forests, in processing and analyzing complex datasets to predict the risk of CVD. These algorithms have shown promising results in identifying patterns and correlations within patient data, which may not beimmediately apparent through traditional clinical assessments. The study also addresses the challenges faced in implementing these algorithms in clinical practice, such as the need for a deep understanding of statistical and clinical knowledge among practitioners and the difficulty in selecting the optimal algorithm for specific research questions or clinical datasets. Despitethese challenges, the research emphasizes the importance of systematic reviews and meta-analyses in evaluating the performance of machine learningalgorithms in heart disease prediction. This approach has been instrumental in identifying the most effective algorithms for clinical application, thereby improving the accuracy and reliability of CVD predictions. Moreover, the research highlights the critical role of machine learning in bridging the gap between data analysis and clinical decision-making. By automating the analysis of large datasets, machine learning models can provide healthcare professionals with valuable insights that may lead to earlier interventions and potentially prevent the onset of CVD.

Looking ahead, the research suggests that futureenhancements in machine learning models for CVD prediction

could focus on incorporating more complex data types, such as genomic data, and improving the interpretability of these models. Additionally, there is aneed for further research to establish acceptable cutoffsfor discrimination measures, such as the area under the ROC curve (AUC), for clinical practice. This will ensure that machine learning models can be effectivelyintegrated into clinical workflows, enhancing the accuracy and efficiency of CVD detection and management. In summary, the use of machine learning models to predict heart disease represents a promising avenue for early detection and prevention strategies. By leveraging the power of data analysis, these models have the potential to significantly improve health outcomes and reduce the burden of CVD on the healthcare system.

References

[1] A. H. M. S. U. Maria Sultana, "Analysis ofData Mining Techniques for Heart Disease rediction," 2018.

[2] M. I. K., A. I., S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

[3] Senthil Kumar Mohan, Chandrasekar Thirumalai, and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" [Access 2019].

[4] S. Kumar, R. Saxena, and S. Mehta, "An Extensive Review on Swarm Robotics," [pp. 140–145, 2009].

[5] B.L Deekshatulua Priti Chandra"Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm" International Conference on Computational Intelligence: Modelling Techniques and Applications [(CIMTA) 2013].

[6] Patel, J., Upadhyay, P., and Patel, "Heart Disease Prediction Using Machine Learning and Data Mining Technique" Journals of Computer Science & Electronics, 2016.

[7] Chavan Patil, A.B. and Sonawane, P. "To Predict Heart Disease Risk and Medications UsingData Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients" International Journal on Emerging Trends inTechnology, 2017.

[8] V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," vol. 38, no. 3, pp. 124–128, 2016.

[9] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," J. Theor. Appl. Inf. Technol., 2009.

[10] Ashish Sharma, Dinesh Bhuriya, UpendraSingh, "Survey of Stock Market Prediction Using Machine Learning Approach", Electronics,Communication and Aerospace Technology (ICECA),2017 International conference of IEEE, 20-22 April2017, pp.1-5.

[11] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2013.

[12] M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol.,vol. 10, pp. 85–94, 2013.

[13] Jaymin Patel, Prof. Tejal Upadhyay, and Dr. Samir Patel, Sep 2015-Mar 2016, "Heart Disease Prediction using Machine Learning and Data Mining Technique", Vol. 7, No.1, pp. 129-137.

[14] Vikas Chaurasia, and Saurabh Pal, 2013, "Early Prediction of Heart Diseases Using Datamining Techniques", Caribbean Journal of Science and Technology, ISSN: 0799-3757, Vol.1, pp.208-218.

[15] Gunsai Pooja Dineshgar, and Mrs. Lolita Singh, February 2016, "A Review on Data Mining for Heart Disease Prediction", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Vol. 5,Issue 2, pp. 462-466.

[16] S. Florence, N. G. Bhuvaneswari Amma, G. Annapoorani, and K. Malathi, November 2014, "Predicting the Risk of Heart Attacks using Neural Network and Decision Tree", International Journal of Innovative Research in Computer and CommunicationEngineering, ISSN (Online): 2320-9801, Vol. 2, Issue11, pp. 7025-7028.

[17] Noura Ajam, 2015, "Heart Diseases Diagnoses Using Artificial Neural Network", Network and Complex Systems, ISSN: 2224-610X(Paper), ISSN: 2225-0603(Online), Vol.5, No.4, pp. 7-11.

[18] Dhanashree S. Medhekar, Mayur P. Bote, andShruti D. Deshmukh, March 2013, "Heart Disease Prediction System Using Naive Bayes", International Journal of Enhanced Research in Science Technology & Engineering, ISSN No: 2319-7463, Vol. 2, Issue 3, pp. 1-5.

[19] Boshra Bahrami, and Mirsaeid Hosseini Shirvani, February 2015, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journal of Multidisciplinary Engineering Science and Technology(JMEST), ISSN:3159- 0040, Vol. 2, Issue 2, pp. 164-168.

[20] Shadab Adam Pattekari, and Asma Parveen, 2012, "Prediction System for Heart Disease using Naive Bayes", International Journal of Advanced Computer and Mathematical Sciences, ISSN: 2230- 9624, Vol. 3, Issue 3, pp. 290-294

[21] Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir, and Y. K. Sharma,19 March 2016, "Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Advent Technology, E-ISSN:2321-9637, SpecialIssue National Conference "NCPC-2016", pp. 104-106.

[22] G. Purusothaman, and P. Krishnakumari, June 2015, "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", Indian Journal of Science andTechnology, Vol.8(12), DOI:10.17485/ijst/2015/v8i12/58385, pp. 1-5.

[23] P.K. Anooj, ―Clinical decision supportsystem: Risk level prediction of heart disease using weighted fuzzy rules; Journal of King Saud University– Computer and Information Sciences (2012) 24, 27– 40. Computer

Science & Information Technology (CS& IT) 59.

[24] Nidhi Bhatla, Kiran Jyoti "An Analysis of Heart Disease Prediction using Different Data Mining Techniques". International Journal of Engineering Research & Technology.

[25] Hollan, J., Hutchins, E. and Kirsh, D.Distributed cognition: Toward a new foundation for human-computer interaction research. ACM TOCHI, 7(2), (2000).

[26] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar,―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[27] Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

[28] Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975– 888).

[29] Erickson, T., Smith, D., Kellogg, W., Laff, M., Richards, J and Bradner, E. Socially translucent conversations: Social proxies, persistent conversation,and the design of "Babble." Proc. ACM CHI (1999). Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

[30] Heon Gyu Lee, Ki Yong Noh, and Keun Ho Ryu, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, 2007.

[31] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Intelligent Computing in Signal Processing and Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[32] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[33] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008.

[34] W.J. Frawley and G. Piatetsky-Shapiro,"Knowledge Discovery in Databases: An Overview", AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.

[35] William Carroll; G. Edward Miller, "Disease among Elderly Americans: Estimates for the US civilian noninstitutionalized population, 2010," Med. Expend. Panel Surv., no. June, pp. 1–8, 2013.