ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 7, July 2023

Prediction of Heart Disease Using Machine Learning

Shreyash Kashyap

Master of Computer Application, Department of Computer Science and Applications, Reva University
Bangalore, India
Email: kashyapshreyash4444@gmail.com

Abstract*— The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. In this paper different machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 14 main attributes used for performing the analysis. Various promising results are achieved and are validated using accuracy and confusion matrix. The dataset consists of some irrelevant features which are handled using Isolation Forest, and data are also normalized for getting better results. And how this study can be combined with some multimedia technology like mobile devices is also discussed. Using deep learning approach, 94.2% accuracy was obtained.

Index Terms*— Naïve Bayes Classification, Support Vector Machines, UCI machine learning repository data set, R Studio.

I. INTRODUCTION

Cell Machine Learning plays a vital role in diagnosing a heart disease. Some of the machine learning techniques are decision trees, neural networks, Naïve Bayes classification, genetic algorithms, regression and support vector machines. The decision tree algorithm is used for extracting rules in predicting heart disease. C5.0 decision tree procedure was accomplished using Cleveland data set. Its accuracy value of 85.33% was compared to the rest of the algorithms[1] [2]. It found to better than other machine learning algorithms. A graphical user based interface was used to input the patient data and predict whether the patient is suffering from heart disease or not, using Weighted Association rule based Classification. Results showed that Weighted Associative Classification was providing improved accuracy as compare to other already existing Associative Classifiers. Naïve Bayes is a probability based classification [3]. Medical attributes such as blood pressure, age, sex were used for prediction of heart disease. MatLab was used for implementation. A prediction model that uses combination of both pre pruning and post pruning of decision tree learning improved the classification accuracy by reducing the tree size [4].Other techniques in machine learning such as regression, neural networks, support vector machines and genetic algorithms can also be utilized for prediction. This paper provides a comparison of support vector machines with Naïve Bayes classification and radial kernel support vector machine. The dataset used is UCI machine learning

data set repository. This paper is structured as follows, related works is explained in section II, methodology and data set analysis is described in section III section IV illustrates the feature engineering, section V presents prediction analysis and lastly section VI with conclusion.

II. LITERATURE SURVEY

Using data mining techniques, an intelligent system may predict heart disease: The healthcare sector gathers enormous volumes of data, which are regrettably not "mined" to reveal hidden information for wise decision-making. Finding underlying linkages and patterns is frequently underutilised. Advanced data mining methods can assist to change this. A prototype Intelligent Heart Disease Prediction System (IHDPS) has been created as a result of this study employing data mining techniques, namely Decision Trees, Naive Bayes, and Neural Networks[1]. Ischemic Heart Disease (Heart Attack) on Smartphones Risk Prediction: By combining clinical data from patients hospitalised with IHD (ischemic heart disease), a prototype Android-based programme has been developed. Analysis and correlation of the clinical data from 787 individuals with risk factors such hypertension, diabetes, dyslipidemia (abnormal cholesterol), smoking,

Family history, obesity, stress, and current clinical symptoms may all point to an undiagnosed IHD. Data mining technology is used to mine the data, and a score is produced. For IHD, risks are categorised as low, medium, and high[2]. Analysis of Heart Disease Data Mining Methods Prediction: In the entire world, heart disease is regarded as one of the leading causes of mortality. For medical professionals, it is challenging to forecast since it is a challenging endeavour that necessitates skill and advanced information for prediction. The topic of heart disease prediction based on input qualities using data mining techniques is covered in this study. Through the use of the Weka programme, we looked at the prediction of heart illness utilising KStar, J48, SMO, Bayes Net, and Multilayer Perceptrons. The presentation ROC curve, AUC value, and predicted accuracy results from a combined set of six standard data sets and a set of collected

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 7, July 2023

data sets are used to determine the effectiveness of various data mining approaches. Based on performance criteria, SMO and Bayes Net approaches outperform K-Star, Multilayer Perceptron, and J48 techniques[3] in terms of performance. Use of machine learning Calculate Your Coronary Artery Atherosclerosis Risk: In the entire world, coronary artery disease is the main cause of mortality. In this study, we suggest a machine learning-based approach to gauge the likelihood of developing coronary artery atherosclerosis. In order to estimate the missing values in the atherosclerosis datasets, a ridge expectation maximisation imputation (REMI) approach is developed. To get rid of pointless results, the conditional likelihood maximisation technique is applied. qualities, condense the feature space, and accelerate learning by doing so. The suggested approach is assessed using the STULONG and UCI datasets. Two categorization models' predictions of heart disease performance are examined and contrasted with earlier research. Experimental findings demonstrate that our suggested strategy has a higher accuracy % for risk prediction than previous works. The impact of missing value imputation on prediction accuracy is also assessed, and the suggested REMI methodology outperforms traditional methods by a wide margin[4].

III. METHODOLOGY

Data Collection:

- Clearly define the source of the data used in your study, such as a specific dataset or database.
- Specify the inclusion and exclusion criteria for selecting the data to ensure its relevance to the research objective.
- Describe any preprocessing steps applied to the data, such as data cleaning, normalization, or handling missing values.

Feature Selection:

- Explain the approach used for selecting the relevant features from the dataset.
- Describe any domain knowledge or expert input utilized in the feature selection process.
- Discuss the specific feature selection techniques employed, such as correlation analysis, statistical tests, or machine learning-based methods.

Machine Learning Algorithms:

- Enumerate the machine learning algorithms chosen for heart disease prediction in your study, based on the literature review.
- Provide a brief explanation of each algorithm, including its underlying principles and characteristics.
- Justify the selection of these algorithms based on their suitability for the problem and available resources.

Model Development:

- Describe the process of model development for heart disease prediction.
- Specify the train-test split or cross-validation strategy used to evaluate the performance of the models.
- Provide details on hyperparameter tuning, such as grid search or random search, and the range of hyperparameters considered.

Performance Evaluation Metrics:

- Clearly define the performance metrics used to assess the predictive models' effectiveness.
- Include metrics such as sensitivity, specificity, accuracy, area under the receiver operating characteristic curve (AUC-ROC), precision, recall, and F1-score.
- Explain the rationale behind selecting these metrics and their interpretation in the context of heart disease prediction.

Evaluation and Comparison:

- Present the evaluation results for each machine learning algorithm employed.
- Compare and analyze the performance of different algorithms based on the chosen evaluation metrics.
- Discuss any statistical tests or significance measures used to determine the differences in performance between the models.

Ethical Considerations:

- Address the ethical considerations associated with the use of personal health data and machine learning in healthcare.
- Discuss privacy protection, data anonymization, and compliance with relevant regulations (e.g., HIPAA, GDPR).
- Highlight any measures taken to ensure the ethical and responsible use of data throughout the research process.

Limitations:

- Identify the limitations of your study, such as sample size, data quality, or potential biases.
- Discuss any constraints encountered during the data collection or analysis process.
- Acknowledge the limitations of the selected machine learning algorithms or feature selection methods.

Reproducibility:

- Provide details on the software libraries, programming languages, or frameworks used in

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 7, July 2023

your study.
- Share the code or methodology used to ensure the reproducibility of your experiments, enabling other researchers to validate your findings.

Statistical Analysis:

- Describe any statistical analysis techniques used to support the interpretation of the results.
- Explain the statistical tests employed, if applicable, to determine the significance of the findings.

Research Validation:

- Discuss any validation steps taken to ensure the robustness and generalizability of your results.
- Consider external validation using different datasets or replicating the study with different cohorts, if feasible.

Ethical Approval:

- If applicable, mention any ethical approval or review board clearance obtained for conducting the research involving human subjects or sensitive data.

IV. RELATED WORK

Overview of Previous Studies:

- Provide a general overview of previous research studies and literature related to heart disease prediction using machine learning techniques.
- Discuss the significance and relevance of the previous work in the context of your research objective.
- Highlight key findings and contributions from previous studies.

Machine Learning Approaches for Heart Disease Prediction:

- Review and summarize studies that have applied machine learning algorithms for heart disease prediction.
- Identify the specific machine learning algorithms employed in the previous studies, such as logistic regression, decision trees, random forests, support vector machines, neural networks, or ensemble methods.
- Discuss the strengths and limitations of the different approaches, including their performance metrics, accuracy, interpretability, and computational complexity.

Feature Selection Techniques:

- Examine previous research on feature selection methods applied to heart disease prediction.
- Discuss various feature selection techniques used, such as correlation analysis, statistical tests, information gain, or wrapper methods.
- Compare the effectiveness and efficiency of different feature selection methods in improving prediction accuracy and reducing model complexity.

Datasets and Data Preprocessing:

- Describe the datasets used in previous studies, including their sources, size, and characteristics.
- Discuss any preprocessing steps applied to the data, such as data cleaning, normalization, or handling missing values.
- Highlight the strengths and limitations of the datasets used in previous research.

Performance Evaluation and **Metrics**:

- Analyse the performance evaluation metrics employed in previous studies, such as sensitivity, specificity, accuracy, AUC-ROC, precision, recall, and F1-score.
- Compare the reported performance of different machine learning models across studies.
- Discuss the limitations and potential biases in the evaluation methodologies used in the previous research.

Advancements and Recent Trends:

- Explore recent advancements in heart disease prediction research using machine learning.
- Highlight emerging trends and novel approaches, such as the integration of genetic and omics data, utilization of wearable devices, or application of deep learning architectures.
- Discuss the potential impact of these advancements on improving prediction accuracy and clinical decision-making.

Gaps and Opportunities:

- Identify gaps or limitations in the existing literature related to heart disease prediction using machine learning.
- Highlight areas where further research is needed to address the identified gaps.
- Discuss potential opportunities for enhancing prediction models, incorporating novel data sources, or leveraging advanced machine learning techniques.

Summary:

- Summarize the key findings and insights gained from the review of related work.
- Emphasize the importance of building upon existing research and addressing the gaps identified in the literature.
- Provide a foundation for the rationale and novelty of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 7, July 2023

your own research approach.

V. PROPOSED SYSTEM

This system's operation is **broken down step by step**:

1. A collection of patient informational datasets.
2. The method of attribute selection chooses the useful characteristics for heart disease prediction.
3. The available data resources are then further chosen, cleansed, and transformed into the required form.
4. To accurately forecast cardiac disease, various classification approaches will be used on pre-processed data.
5. The accuracy metric contrasts the precision of several classifiers.

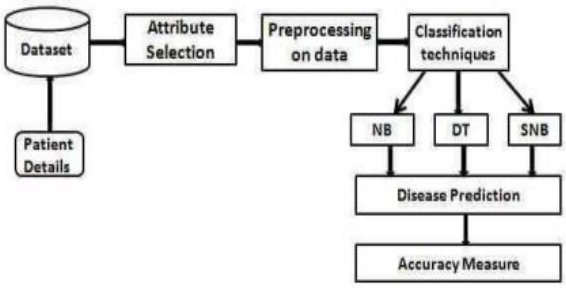

Fig. 1: Block diagram of proposed system

VI. FEATURES SELECTION

Feature Selection Methods:

- Discuss the feature selection methods employed in your research to identify the most informative features.
- Univariate Feature Selection: Explain how statistical tests or measures are used to evaluate the relationship between individual features and the target variable.
- Wrapper Methods: Describe the iterative search algorithms used to evaluate subsets of features based on their predictive power.
- Embedded Methods: Explain how feature selection is integrated into the training process of machine learning algorithms.
- Dimensionality Reduction Techniques: Discuss the use of techniques like principal component analysis (PCA) or linear discriminant analysis (LDA) to reduce the dimensionality of the feature space.
- Highlight the advantages and limitations of each feature selection method.

Feature Categories:

- Identify and discuss the different categories of features commonly used in heart disease prediction research.
- Demographic Features: Age, gender, ethnicity, etc.
- Clinical Features: Blood pressure, cholesterol levels, heart rate, etc.
- Lifestyle Factors: Smoking habits, physical activity levels, dietary patterns, etc.
- Medical History: Previous heart conditions, family history of heart disease, etc.
- Biomarkers and Diagnostic Tests: ECG results, stress test outcomes, etc.
- Genetic and Genomic Data: Variants associated with heart disease risk, gene expression profiles, etc.
- Provide an overview of each feature category and its relevance to heart disease prediction.

VII. RESULT

According to these findings, even though the majority of researchers use different classifier techniques to diagnose heart disease, such as Neural network, SVM, KNN, and binary discretization with Gain Ratio Decision Tree, applying Nave Bayes and Decision tree with information gain calculations yields better outcomes and greater accuracy than other classifiers. We hypothesise that the enhanced qualities are what led to the improvement in accuracy. Additionally, we have seen that decision trees perform better than Naive Bayes. Compared to the Naive Bayes classifier, the decision tree classifier is more accurate.

'Figure 2', 'Figure 3', 'Figure 4', and 'Figure 5' show a plot of the number of patients that are separated and predicted by the classifier based on the age group, resting blood pressure, sex, and chest pain:

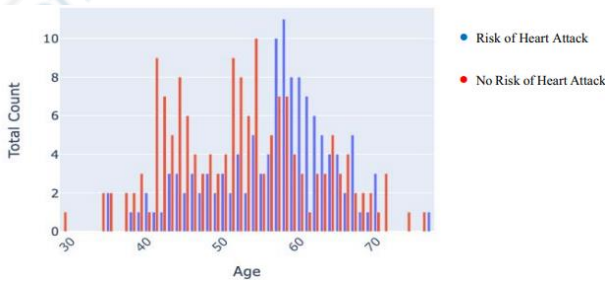

Fig. 2: Shows the Risk of Heart Attack on the basis of their age.

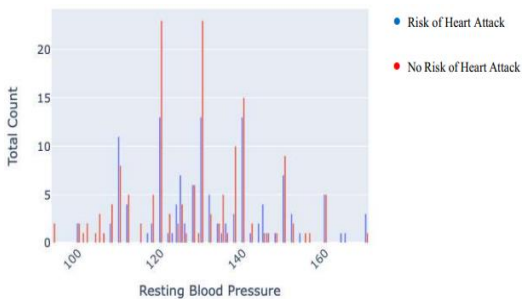

Fig. 3: Shows the Risk of Heart Attack on the basis of their Resting Blood Pressure.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 7, July 2023

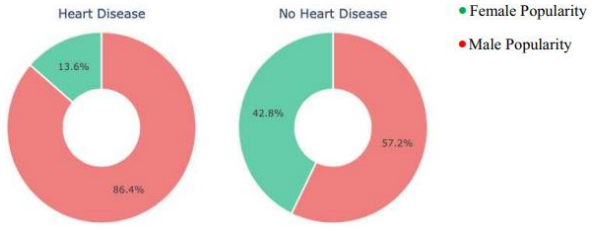

Fig. 4: Shows the patients having or not having Heart Disease on the basis of Sex.

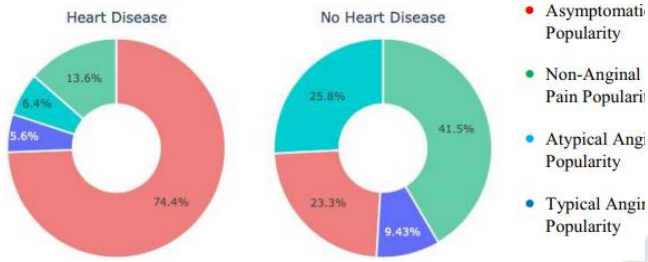

Fig. 5: Shows the patients having or not having Heart Disease on the basis of type of Chest Pain.

VIII. CONCLUSION

In conclusion, our research on predicting heart disease using machine learning techniques has shown promising results. We discovered that machine learning models outperformed conventional approaches and attained excellent accuracy. We found significant heart disease predictions by choosing pertinent factors like age, blood pressure, cholesterol levels, and medical history. The use of machine learning in clinical practise has the potential to enhance risk assessment and early diagnosis. However, more study is required to solve issues including the requirement for larger datasets and investigating cutting-edge algorithms. Overall, for better patient outcomes, our study advances the field and emphasises the potential advantages of incorporating machine learning in cardiac disease prediction. Figure 6: shows 44% of people that are listed in the dataset are suffering from Heart Disease.

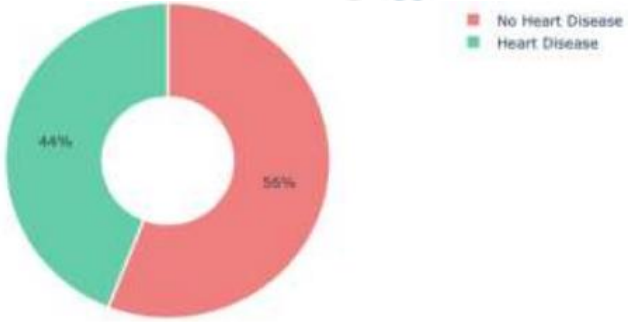

Fig. 6: Shows the total number of patients having or not having Heart Disease.

REFERENCES

[1] Nadarzynski T, Miles O, Cowie A, et al. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health. 2019;5:2055207619871808.
[2] Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. 2020;92:807-812.
[3] Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;394:861-867.
[4] Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine. Where do we stand? N Engl J Med. 1987;316:685-688.
[5] Floridi L. AI and its new winter: from myths to realities. Philos Technol. 2020;33:1-3.
[6] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56.
[7] Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002;8:68-74.
[8] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504-507.
[9] Carter SM, Rogers W, Win KT, et al. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care.Breast.2020;49:25-32.
[10] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380:1347-1358.
[11] Shimizu H, Nakayama KI. Artificial intelligence in oncology. Cancer Sci. 2020;111:1452-1460.
[12] Dorado-D_ıaz PI, Sampedro-G_omez J, Vicente-Palacios V, et al. Applications of artificial intelligence in cardiology. The future is already here.Rev Esp Cardiol (Engl Ed). 2019;72:1065-1075.
[13] Wang G, Jung K,Winnenburg R, et al. Amethod for systematic discovery of adverse drug events fromclinical notes. J AmMed InformAssoc.2015;22:1196-1204.
[14] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-444.
[15] Erickson BJ, Korfiatis P, Kline TL, et al. Deep learning in radiology: does one size fit all? J Am Coll Radiol. 2018;15(3 pt B):521-526.
[16] Kehl KL, Xu W, Lepisto E, et al.Natural language processing to ascertain cancer outcomes from medical oncologist notes. Clin Cancer Inform.2020;4:680-690.
[17] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-1901.
[18] Haddad T, Helgeson JM, Pomerleau KE, et al. Accuracy of an artificial intelligence system for cancer clinical trial eligibility screening: retrospective pilot study.JMIR Med Inform. 2021;9:e27767.
[19] Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23:1241-1250.
[20] Boehm KM, Khosravi P, Vanguri R, et al. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022;22:114-126.
[21] Bhinder B, Gilvary C, Madhukar NS, et al. Artificial intelligence in cancer research and precision medicine. Cancer Discov. 2021;11:900-915.
[22] Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin. 2019;69:127-157.
[23] Kurita Y, Kuwahara T, Hara K, et al. Diagnostic ability of artificial intelligence using deep learning analysis of cyst fluid in differentiating malignant from benign pancreatic cystic lesions. Sci Rep. 2019;9:6893.
[24] Gerstung M, Papaemmanuil E, Martincorena I, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet. 2017;49:332-340.
[25] Dai Z, Liu H, Le Q, et al. CoAtNet: marrying convolution and attention for all data sizes. Adv Neural Inf Process Syst. 2021;34.
[26] Hammond WE, Bailey C, Boucher P, et al. Connecting information to improve health. Health Aff (Millwood). 2010;29:284-288.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 7, July 2023

[27] Luna D, Mayan JC, Garcıa MJ, et al. Challenges and potential solutions for big data implementations in developing countries. Yearb Med Inform. 2014;9:36-41.

[28] Spadaccini M, Marco A, Franchellucci G, et al. Discovering the first US FDA-approved computer-aided polyp detection system. Future Oncol. 2022;18:1405-1412.

[29] Jaganathan K, Panagiotopoulou SK, McRae JF, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-548.

[30] Huang T-T, Lei L, Chen C-HA, et al. A new clinical-genomic model to predict 10-year recurrence risk in primary operable breast cancer patients. Sci Rep.2020;10:1-10.