International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 3 (2025)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.org
Vol. 21, No. 3, 2025

Research Paper

HEART DISEASE IDENTIFICATION ON METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALT CARE

P.Narendra *1, P. Adhi lakshmi*2

*1 Research Scholar, Department Of CSE, NIE, JNTUK, Andhra Pradesh

*2 Associate Professor, Department Of CSE, NIE, JNTUK, Andhra Pradesh

Received: 07-5-2025 Accepted: 14-6-2025 Published: 25-6-2025

ABSTRACT

Heart disease remains one of the leading causes of mortality worldwide, making early detection and timely intervention critical. This project, titled *"Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare"*, aims to develop an intelligent, cost-effective cardiology solution to assist in the clinical decision-making process. Leveraging the power of machine learning, the system is designed to predict whether a patient is at risk of heart disease based on various medical parameters. The model performs binary classification where the outcome is either positive (1) — indicating the presence of heart disease — or negative (0) — indicating its absence. The project ensures that healthcare providers can access a reliable tool for diagnosing heart conditions, ultimately improving patient outcomes and optimizing healthcare resources. By integrating this model into electronic healthcare systems, it supports scalable, data-driven, and clinically appropriate care for patients with suspected cardiovascular conditions.

Keywords: Heart disease prediction, Machine learning, Logistic Regression, KNN, Random Forest, Feature selection, E-healthcare, Binary classification, Clinical decision support system.

1.INTRODUCTION

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world [1]. The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen [2]. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much

effective in early time identification due to several reasons, such as accuracy and execution time [3]. The diagnosis and treatment of heart disease is extremely difcult when modern technology and medical experts are not available [4]. The effective diagnosis and proper treatment can save the lives of many people [5]. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually [6]. Most of the people in the United States are suffering from heart disease [7]. Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difcult to analyze [8].

Thus, to develop a noninvasive diagnosis system based on classiers of machine learning (ML) to resolve these issues. Expert decision system based on machine learning classiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a heart disease data set was used by various

researchers [11] and [12] for the identification problem of HD. The machine learning predictive models need proper data for training and testing. The performance of machine learning model can be increased if balanced dataset is use for training and testing of the model. Furthermore, the model predictive capabilities can improved by using proper and related features from the data. Therefore, data balancing and feature selection is significantly important for model performance improvement. In literature various diagnosis techniques have been proposed by various researchers, however these techniques are not effectively diagnosis HD. In order to improve the predictive capability of machine learning model data preprocessing is important for data standardization. Various Preprocessing techniques such removal of missing feature value instances from the dataset, Standard Scalar (SS), Min-Max Scalar etc.

The feature extraction and selection techniques are also improve model performance. Various feature selection techniques are mostly used for important feature selection such as, Least-absolute-shrinkage-selection-operator (LASSO), Relief, Minimal-Redundancy-Maximal-Relevance (MRMR), Local-learning-based features-selection (LLBFS), Principle component Analysis (PCA), Greedy Algorithm (GA), and optimization methods, such as Anty Conley Optimization (ACO), fruit y optimization (FFO), Bacterial Foraging Optimization (BFO) etc. Similarly Yun et al. [13] presented different techniques for different type of feature selection, such as feature selection for high-dimensional small sample size data, large-scale data, and secure feature selection. They also discussed some important topics for feature selection have emerged, such as stable feature selection, multi view feature selection, distributed feature selection, multi-label feature selection, online feature selection, and adversarial feature selection. Jundong et al. [14] discussed the challenges of feature selection (FS) for big data. It is necessary to decrease the dimensionality of data for various learning tasks due to the curse of dimensionality.

Feature selection has great inuence in numerous applications such as building simpler, increasing learning performance, creating clean and understandable data. The feature selection from big data is challenging job and create big problems because big data has many dimensions. Further, challenges of feature selection for structured, heterogeneous and streaming data as well as its scalability and stability issues. For big data analytics challenges of feature selection is very important to resolved. In [15] designed unsupervised hashing scheme, called topic hyper graph hashing, to report the limitations. Topic hypergraph hashing effectively mitigates the semantic shortage of hashing codes by exploiting auxiliary texts around images. The proposed Topic hyper graph hashing can achieve superior performance equaled with numerous state-of-theart approaches, and it is more appropriate for mobile image retrieval.

The feature selection algorithms are classified into three type such as lter based, wrapper based and embedded based. All these feature selection mechanisms have some advantages and limitations in certain cases. The lter based method measures the relevance of a feature by correlation with the dependent variable while the wrapper feature selection algorithm measure the usefulness of a subset of features by actually training the classier on it. The lter method is less computationally complex than wrapper method. The feature set selected by the lter is general and can be applied to any model and it is independent of a specic model. In feature selection global relevance is of greater importance. On another hand suitable machine learning model is necessary for good results. Obviously, a good machine learning model is a model that not only performs well on data seen

during training (else a machine learning model could simply learn the training data), but also on unseen data. To evaluate all classiers on data and that they get, on average, 50% of the cases right [16]. Furthermore, appropriate cross validation techniques and performance evaluation metrics are critical necessary for a model when model is train and test on dataset. We proposed a machine learning based diagnosis method for the identification of HD in this research work. Machine learning predictive models include ANN, LR, K-NN, SVM, DT, and NB are used for the identification of HD. The standard state of the art features selection algorithms, such as Relief, m RMR, LASSO and Local-learning-based features- selection (LLBFS) have been used to select the features We also proposed fast conditional mutual information (FCMIM) features selection algorithm for features selection. Leave-one-subject-out cross-validation (LOSO) technique has been applied to select the best hyper-parameters for best model selection. Apart from this, different performance assessment metrics have been used for classiers performances evaluation. The proposed method has been tested on Cleveland HD dataset.

Furthermore, the performance of the proposed technique have been compared with state of the art existing methods in the literature, such as NB [17], Three phase ANN (Articial neural Network) diagnosis system [18], Neural network ensembles (NNE) [19], ANN-Fuzzy-AHP diagnosis system (AFP) [20], Adaptive weighted-Fuzzy-system-ensemble (AWFSE) [21]. The research study has the following contributions. Firstly, the authors try to address the problem of features selection by employing pre-processing techniques and standard state of the art four features selection algorithms such as Relief, mRMR, LASSO, and LLBFS for appropriate subset of features and then applied these features for effective training and testing of the classiers that identify which feature selection algorithm and classier gives good results in term of accuracy and computation time. Secondly, the authors proposed fast conditional mutual information (FCMIM) FS algorithm for feature selection and then these features are input to classiers for improving prediction accuracy and reducing computation time. The classiers performances have been compared on features selected by the standard state of the art FS algorithms with the selected features of the proposed FS algorithm. Thirdly, identify weak features from the dataset which affect the performance of the classiers. Finally, suggests that heart disease identification system (FCMIM-SVM) effectively identify the HD. The paper remaining sections are structured as follows. The literature related to the problem has been discussed in section 2. In section 3 the dataset and the theoretical and mathematical knowledge of feature selection and classification algorithms are discussed in details. Additionally, discuss the technique of cross-validation and performance measuring metrics. In section 4 results of all experiments are analyzed and discussed in details. The last section 5 the conclusion and future direction of the research work have been explored in details.

2.METHODOLOGY

The methodology followed in this project comprises a sequence of well-structured steps including data preprocessing, feature selection, algorithm implementation, model evaluation, and web deployment. The focus is on accurately predicting the presence or absence of heart disease using machine learning classification algorithms.

1. Data Collection and Preprocessing:

The first step involves acquiring a heart disease dataset, typically from a reliable source like the UCI Machine Learning Repository or Kaggle. The dataset consists of multiple features such as age, sex, resting blood pressure, cholesterol level, fasting

blood sugar, maximum heart rate, and other clinical parameters.

Data preprocessing includes handling missing values, encoding categorical variables, and normalizing numerical data. Normalization is essential for algorithms like KNN that rely on distance metrics. We use Min-Max normalization defined as:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where X is the original value, and X' is the normalized value ranging between 0 and 1.

2. Feature Selection and Outlier Detection:

In the second approach, we perform **feature selection** to reduce dimensionality and improve model performance. Techniques like correlation matrix analysis and Recursive Feature Elimination (RFE) are used to identify the most impactful features.

Outliers are detected using statistical methods such as the **Z-score** and **Interquartile Range (IQR)** method. The IQR method identifies outliers as:

$$IQR = Q3 - Q1$$

$$\text{Lower Bound} = Q1 - 1.5 \times IQR, \quad \text{Upper Bound} = Q3 + 1.5 \times IQR$$

Values outside these bounds are considered outliers and can be removed or treated.

3. Model Building:

Three classification algorithms were implemented:

Logistic Regression: This is a statistical model that predicts the probability of a binary outcome using the logistic function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

where YY is the target variable (presence of heart disease), XiX_i are the features, and βi\beta_i are the model coefficients.

K-Nearest Neighbors (KNN): This is a distance-based classifier that assigns the class based on the majority vote of the k-nearest data points. The Euclidean distance used is:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Random Forest: This is an ensemble method based on decision trees. It creates multiple trees on random subsets of data and features and outputs the mode of predictions from individual trees.

4. Model Evaluation:

The models were evaluated based on **accuracy**, **precision**, **recall**, and **F1-score**. The formulas are as follows:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

TP: True Positives

TN: True Negatives

FP: False Positives

FN: False Negatives

Among the three models, Logistic Regression achieved the highest test accuracy and F1-score after normalization and feature selection, making it the most suitable for this binary classification task.

3. RESULT AND DISCUSSION

By applying different machine learning algorithms and then using deep learning to see what difference comes when it is applied to the data, three approaches were used. In the first approach, normal dataset which is acquired is directly used for classification, and in the second approach, the data with feature selection are taken care of and there is no outliers detection. results which are achieved are quite promising and then in the third approach the dataset was

normalized taking care of the outliers and feature selection.results achieved are much better than the previous techniques, and when compared with other research accuracies, our results are quite promising

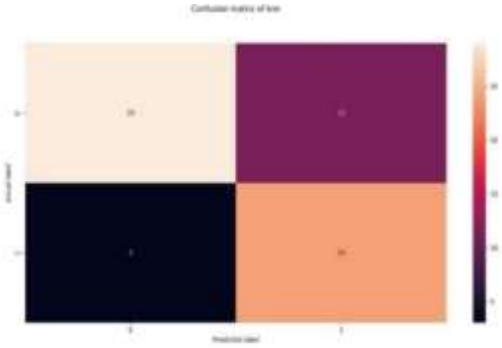

Fig1: Confusion matrix

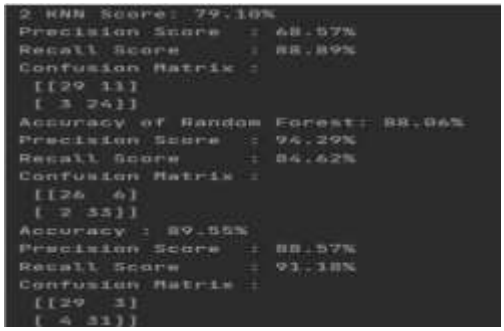

Fig2: Comparison of three algorithms

Here we observe and compare the accuracy of three models namely Logistic Regression, KNN, Random Forest among these, Logistic Regression model has the best overall accuracy and F1 score. Therefore, we should use Logistic Regression algorithm to predict the heart disease.

4.CONCLUSION

Clinical finding is a significant region of exploration which assists with recognizing the event of a coronary illness. The framework, utilizing different methods referenced, will thus uncovered the root coronary illness alongside the arrangement of most plausible heart Diseases which have comparative side effects. The information base utilized is a portrayal data set so to decrease the dataset tokenization, separating and stemming is finished. The venture presents a novel mixture model to recognize and affirm CAD cases requiring little to no effort by utilizing clinical information that can be effectively gathered at clinics. Intricacy of framework is diminished by decreasing the dimensionality of the informational collection with PSO. It gives reproducible and target finding, and subsequently can be a significant extra device in clinical practices. Results are equivalently, encouraging and along these lines the proposed half and half technique will be useful in coronary illness diagnostics. Trial results exhibit the predominance of the proposed half breed technique concerning forecast precision of CAD.

5.FUTURE WORK

In this project three methods in which comparative analysis was done and promising results were achieved. *e conclusion which we found is that machine learning algorithms performed better in this analysis. Many researchers have previously suggested that we should use ML where the dataset is not that large, which is proved in this paper. The methods which are used for comparison are confusion matrix, precision, specificity, sensitivity, and F1 score. For the 13 features which were in the dataset, K-Neighbours classifier performed better in the ML approach when data pre-processing is applied.

6.REFERENCES

1. S. I. Ansarullah and P. Kumar, __A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method,‘‘ Int. J.Recent Technol. Eng., vol. 7, no. 6S, pp. 1009–1015, 2019

2. A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, __Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data,‘‘ Sensors, vol. 20, no. 9, p. 2649, May

2020
3. A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, __Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection,‘‘ in Proc. IEEE 5th Int. Conf. Converg.Technol. (ICT), Mar. 2019, pp. 1–4
4. U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, __A novel integrated diagnosis method for breast cancer detection,‘‘ J. Intell. Fuzzy Syst, vol. 38, no. 2, pp. 2383– 2398, 2020.
5. T. Hastie, R. Tibshirani, and J. Friedman, ―*e elements of statistical learning,‖ Data Mining, Inference, and Prediction, Springer, Cham, Switzerland, 2020.
6. S. Marsland, ―Machine learning,‖ An Algorithmic Perspective, CRC Press, Boca Raton, FL, USA, 2020.
7. P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, ―Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability,‖ IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 3, pp. 727–733, 2013.
8. M. M. A. Rahhal, Y. Bazi, H. Alhichri, N. Alajlan, F. Melgani, and R. R. Yager, Deep learning approach for active classification of electrocardiogram signals,‖ Information Sciences, vol. 345, pp. 340–354, 2016.
9. G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, ―A machine learning system to improve heart failure patient assistance,‖ IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.
10. R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, ―Automatic methods to extract New York heart association classification from clinical notes,‖ in Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1296–1299, IEEE, Kansas City, MO, USA, November 2017.