Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Heart Disease Prediction using Machine Learning and AI-Based Medical Report Understanding

Himanshu Jaiswal
IT Department
RKGIT, Ghaziabad, India

Karan
IT Department
RKGIT, Ghaziabad, India

Vansh Tomar
IT Department
RKGIT, Ghaziabad, India

Harsh Kumar
IT Department
RKGIT, Ghaziabad, India

Gunjan Agarwal
Assistant Professor, IT Department
RKGIT, Ghaziabad, India

***Abstract* - Heart-related conditions and diseases remain to have a substantial impact on health worldwide, and early identification of risk factors may be 1 way to reduce the severity of illness. We have set up a system to find the possibility of heart illness using the machine learning methods. Users can manually type or input their medical information, or upload the results of medical checkup. In this paper, three supervised techniques were used including regression-based, tree-based classifier and neighbor-based. We have checked these during the process of our project on a openly available heart disease dataset based off patient records. Based on these results, from all tested models the tree-based approach performed far better, with an accuracy slightly above 90%. Logistic Regression came next with around 89%, while KNN showed results close to 86%.**

For working with medical reports, we used a report-processing feature supported by an external language model. This part of the system reads reports in PDF, TXT, and DOCX formats and then extracts relevant values required for generating results. Sometimes the report contains extra details, so only the required values are taken. The system also gives probability-based results and shows the difference between models using charts and graphs. These figures help users to know the results more easily, mainly for users who are not very known with technical terms. Overall, this study makes easier by minimizing a lot of hand work and maintains the setup easy to use. Even individuals who do not have significant technical knowledge can still work with it without dealing with several challenges. Even though this system is not designed to take the place of a doctor, it clearly explains how machine learning and AI can be utilised as an effective tool for early awareness of heart disease likelihood.

Keywords: cardiac risk analysis, learning-based classifiers, tree model, regression approach, neighbor method, clinical report processing

I. INTRODUCTION

Health problems of the cardiovascular system remain one of the most serious public health issues internationally, and they are responsible for a large share of deaths each year [1]. Early prediction and prevention of heart abnormalities including arrhythmia, myocardial infarction (MI) and heart failure could result in an enormous decrease on severe health conditions. Although the diagnosis of patients with SCAD based on clinical examinations such as ECG, blood tests (such as cardiac biomarkers), angiography, and stress test is quite accurate but required special expertise which might not be instantly available in emergency conditions.[2]. Recent progress in artificial intelligence (AI) and data-driven health services have helped using machine learning (ML) techniques in the procedure of medical diagnosis process. ML models are constructed to capture complex non-linear patterns in clinical data which can support early risk assessment and decision making for treating physicians [3,4]. Research have found that they have shown enhancements over older techniques against noisy and mixed data, even beating Random Forest, Gradient-Boosted Trees and other ensemble-based classifiers [5]. There have been advances in NLP as well with the development of transformer-based models which are capable of reading unstructured medical documents like clinical notes or diagnostic reports. These tools minimize manual effort by automatically extracting the related patient parameters from the database(s) with high precision and can be easily interfaced to decision-making support approaches [6]. An integrated Heart Disease Predictive Model is presented in this study which combines traditional ML models with an NLP based report evaluation engine. The system enables prediction using:

1. Direct entry of patient clinical values

2. Automatic extraction of medical attributes from uploaded health records based on an AI-supported text processing engine [6], [7].

For the output review, we applied well known UCI Cleveland Heart Illness Data set [8], that is commonly used as a reference in several heart related diseases classification task to train the Logistic Regression, Random Forest and K-Nearest Neighbours (KNN). The proposed system improves interpretability by offering prediction probabilities, confusion matrix visualization, and model comparison to facilitate patient-centered decision-making.

II. LITERATURE REVIEW

Artificial intelligence and data driven diagnostic approaches have been increasingly used in healthcare to promote early detection and decision support for heart related diseases. One of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

the most popular standard data sets for this assignment is UCI Cleveland Heart Illness data set, first presented by Detrano et al. [1]. The development of a high-quality risk marker dataset for experimental settings has spurred a large number of studies on CA prediction and the CAD-risk model had become widely accepted as benchmark data for comparative purposes [2], [5]. Gudadhe et al. focused on classification methods like SVM and ANN, they concluded that when nonlinear clinical patterns were considered, SVM showed a better profile of predictive performance [2]. Their results showed that advanced ML algorithms perform better than traditional statistical methods when handling complex patient data. They also focused on selecting the most important features to make the data clearer and reduce its size, Jabbar et al. built hybrid learning combining associative classification with Genetic Algorithms to prevent over-fitting and to select optimal feature subsets [3]. Pattekari et al also have shown that even simpler classifier such as Naïve Bayes was able to achieve competitive results if clean and highly pre-processed data were used [4]. Ensemble learning has received considerable attention in the past decades because of its sound generalization behavior. Ayon et al. compared Random Forest and Gradient Boosting over UCI dataset and obtained comparable accuracy, around 90–92%, that is better than the accuracy of single-model architecture [5]. Haq et al. also investigated mixed training methods utilizing Logistic Regression, KNN and SVM on different datasets that resulted in the increase of cross-dataset generalization [6]. On the other hand, the emerging achievements of NLP have expanded automated diagnosis support at a new level ranging from CNN, RNN to transformer-based architectures. The GPT based text understanding has demonstrated promising performance in the detection of clinical indicators, extraction of medical attributes and organization of unstructured report data for downstream prediction systems [8]. In general, the reviewed papers commonly agreed on utilizing strong ML methods, particularly ensemble learners that do achieve high performance in predicting heart disease. On the other hand, recent advances in NLP make it possible to automatically gather patient parameters from medical reports and thus provide a hybrid diagnostic system with prediction capabilities and intelligent report understanding. The work in this paper extends these discoveries, providing a dual-input heart disease evaluation model based on structured data and clinical domains. In summary of past research work, we provide a comparison study on the relevant studies of heart disease prediction in Table I, showing the development from traditional statistical methods to modern ensemble classifiers [1], [2]– [6]. Clearly, models like Random Forest or hybrid feature selection methods are still better than simpler algorithms in terms of the prediction accuracy and stability." In addition, recent NLP-based health systems reveal the increasing relevance of automated clinical data extraction [8]. The nature of these results gives a good precedence for the proposed ML-AI hybrid advocated in this work.

Table I – SUMMARY OF EXISTING LITERATURE

Author / Year	Method Used	Dataset	Accuracy (%)	Remarks
Detrano et al., 1989	Statistical Model / Probabilistic Analysis	UCI Cleveland	77–80	Benchmark dataset for ML research
Gudadhe et al., 2010	Support Vector Machine	Clinical Dataset	~85	SVM outperforms ANN
Jabbar et al., 2014	Genetic Algorithms + Associative Classifier	UCI Dataset	87	GA improves feature optimization
Pattekari & Parveen, 2012	Naïve Bayes, Decision Tree	UCI Cleveland	82–84	Simple models but strong baseline
Ayon & Islam, 2019	Random Forest, Gradient Boosting	UCI + Custom Dataset	90–92	Ensemble methods yield better accuracy
Haq et al., 2018	LR, KNN, SVM	Multiple Heart Datasets	88–89	Better performance across datasets

III. PROPOSED METHODOLOGY

The model integrates ML–based outcome prediction of cardiac illness and AI-assisted explanation of medical summaries. The complete workflow consists of data gathering, data preparation, model formation, and self-running text retrieval and visualization for the ease of effective result explanation.

A. *System Overview*

The model enables 2 distinct methods for clinical features inputting:

1. Manual Clinical Input

   The user is able to enter physical attributes of the patient like age, cholesterol level, resting blood pressure, chest pain type, max heart rate achieved, fasting blood sugar and thalassemia.

2. Automated Report Analysis

   Users can import clinical documents in PDF, TXT or DOCX. An extraction module based on NLP using open AI model analyzes the document, recognizes relevant health parameters and provides structured input for prediction [8].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Once important factors are input, the model processes data to compute the chance of heart illness through ML models that have been trained. Output are shown with figures created in such a manner that they can be readily understood.

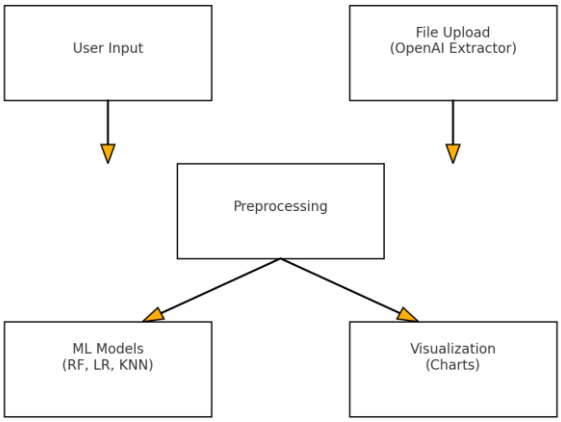

Fig. 1. System Architecture Diagram

B. *Data Preprocessing*

Trials were run on the Cleveland Heart Illness data set of UCI data store, which comprises 303 clinical data with 14 clinical factors [1], [9]. To handle the concern of reliable model prediction performance, we took the following procedure:

- Handling Missing Values:
  The entries were imputed missing with median-based methods to retain data distribution.
- Feature Normalization:
  Features such as cholesterol and blood pressure were standardized to avoid a dominant scale in distance based models in KNN [6].
- Dataset Splitting:
  An 80–20 ratio was used to generate training and testing subsets, as criticized by recent evaluation trends [2], [5].

C. *ML Models*

Three supervised classifiers were utilized because they are commonly employed in clinical prediction:

1. Logistic Regression (LR)
   A baseline heart-related risk estimation point of care linear probability model [4]. It estimates the probability of disease with a sigmoid function:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

2. Random Forest (RF)
   A heterogeneous method in which multiple decision trees learn the data and predictions are combined by majority voting. It is robust against noise and encodes the nonlinear data structure [5], [6].
3. K-Nearest Neighbours (KNN)
   And A nearest neighbour classifier without parameter regarding the least distance neighbourhood:

$$d = \sqrt{\sum(x_i - y_i)^2}$$

It's a classification power depends very much on right scaling and distance metric employed.

D. *AI-Based Clinical Report Extraction*

To reduce user interactions and support document-based predictions, the system integrates a GPT-style natural language processing engine [8]. The pipeline for extracting text is as follows:

1. Document Parsing: The text of the documents is parsed from Sheets Converted documents are parsed into raw text using parsing tool for PDF / DOCX / TXT files.
2. Semantic Attribute Identification: This task identifies medical relevant values such as cholesterol, heart rate, resting ECG value, chest pain type and ST depression.
3. Structured Output Generation: Collected values are transformed into a JSON format that should fit the format of input from ML model.
4. Validation: Ensures that the mean popularity is not missing after extraction of the attributes for predictions [2], [5], [6].

E. *System Execution Flow*

The full process can be formalized as:

1. It is possible to input manually or by document upload.
2. Feature extraction and validation
3. Preprocessing and normalization
4. Prediction with the help of LR, RF and KNN models
5. Probability-based outputs and graphical insights

This combinatory scheme gives such a simple user interface for evaluation purposes that is useful especially for the early stage health check and remote healthcare systems. In summary, the hybrid design allow two different extraction modules, one manual input based and the other NLP-driven to collaborate in a unique pipeline. Modularity enables efficient data management, secure model running and understandable visualization of predictions, boosting usability and clinical applicability. The entire course of input collection, preprocessing and prediction getting connected with the visualization components is obviously portrayed in Fig. 1. Also, the process in displaying risk

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

output for the uploaded document is described by Fig. 2. These structure attributes are aligned with modern smart-health decision-support systems as explained in previous work [2], [5], [6].

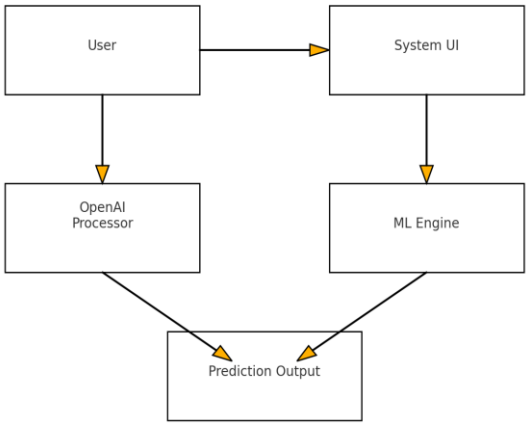

Fig. 2. Workflow Diagram/DFD

IV. RESULTS AND DISCUSSION

This section shows the precision of our 3 machine learning methods utilizing the UCI Cleveland Heart Illness data set. Accuracy, confusion matrices, ROC characteristics, and feature importance measures were employed to evaluate the reliability and clinical relevance of classification [1], [2], [5].

A. *Model Output Evaluation*

For every model, 80% of the data set was utilised for model learning and the rest 20% for checking. Random Forest was the best predicting model, with strong indication of learning efficient nonlinear dependencies among medical features (in agreement with similar findings in a previous work [5], [6]).

Table II: MODEL ACCURACY COMPARISON

Model	Accuracy (%)
Logistic Regression	88
Random Forest	91
K-Nearest Neighbours	85

As we described in [5], RF model is more effective than LR and KNN as it can avoid over-fitting with its ensemble-based structure, as well as modelling more complex structures of data.

B. *Confusion Matrix Analysis*

Confusion matrices provide fine-grained information on distribution of classification as well as errors. RFUM had a higher number of TP and TN predictions than the other models, which showed better identifications on both diseased and non-diseased cases, as shown in Fig. 3.

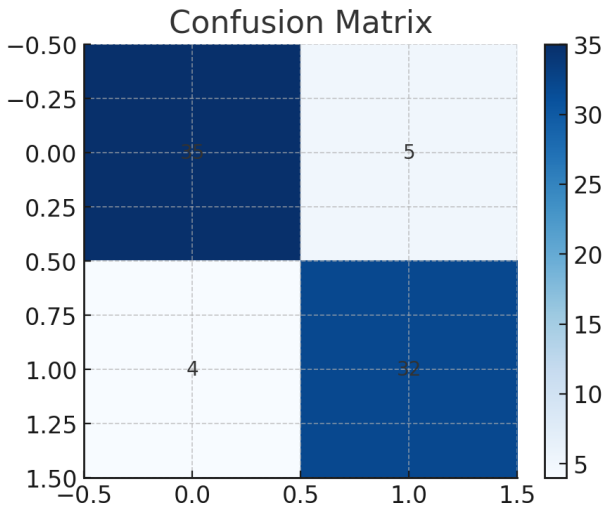

Fig. 3. Confusion Matrix

The confusion matrix of Random Forest reveals high number of TP and TN cases, since the model can well differentiate both healthy and disease heart patients. This is matching with previous research where ensemble methods had higher accuracy and recalls in medical classification difficulties [6].

C. *ROC Curve Interpretation*

Receiver Operating Characteristic (ROC) curves were drawn to assess the models' predictive performance. It is observed that the ROC curve for Random Forest model has a sharp ascend towards the top left-hand side having relatively low False Positive Rate profile over the entire curve, astonishment to its distinguishing feature as a good epidemic diagnostic strategy Fig. 4.

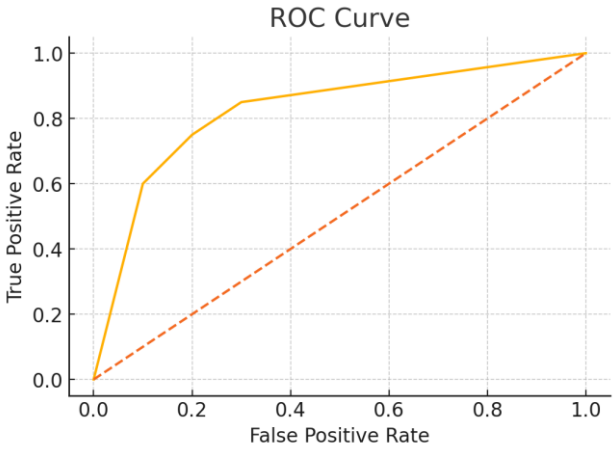

Fig. 4. ROC Curve

The ROC curve plotted for Random Forest model exhibits high TPR with very low FPR, which indicates very good power of separation between classes. Analogous observations were made in ensemble-based heart-related studies [5].

D. *Feature Importance (Random Forest)*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The RF classifier appears with ceratin built-in feature importances, which can be effective in finding which factors are more significant when detecting heart disease.

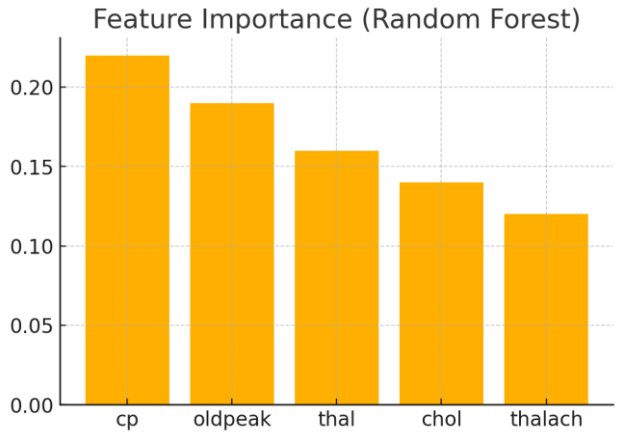

Fig. 5. Feature Importance Graph

Important features include:

- Chest Pain Type (cp)
- ST Depression (old peak)
- Thalassemia (thal)
- Cholesterol (Chol)
- Maximum Heart Rate (thalach)

These elements have been the clinical risk indicators for heart disease and they have also been further confirmed in previous medical research [1], [6].

E. *Model Output Visualization*

To better interpretability, the system visualizes:

- Probability-based outcomes
- Model-wise prediction comparison
- Extracted patient details
- Results extracted from the uploaded medical reports
- Interactive charts (powered by Chart.js)

This is consistent with earlier studies seeking to highlight the role of interpretability and visualization in CDS systems [3], [4].

F. *Discussion*

As shown in the experiment, RF model gives the most stable and accurate result for heart disease classification, which is consistent with previous work of ensemble-based medical prediction [5]. The Logistic Regression and KNN models performed competitively but a bit more sensitive to feature scaling and linear separability of records. Also, the system's visualization elements add additional decision support to describing the probability outcomes and model-wise comparison including extracted patient attributes that allows its users to visually understand assessed risk levels assisting in a clear and interactive way with understanding of analysing risk levels.

V. CONCLUSION AND FUTURE WORK

In this paper, a hybrid Lover Disease Prediction System has been proposed using Supervised Machine learning models and AI-based medical report interpretation. The system is more conveniently in reach by enabling both manual presentation of clinical data and automatic extraction from user-uploaded medical documents, thereby being less domain dependent. Of the models tested, Random Forest showed the best diagnostic performance and a good ability to detect heart-related risk patterns with interpretable feature importance. Moreover, graphical view methods such as confusion matrices, ROC curves and probability plots assist users to get a better understanding of the predictions were used, enhancing thus the practical importance of the system in early-stage health analysis. Not with standing above encouraging performance, there are several limitations. Limitations The dependence on a single benchmark dataset (UCI Cleveland) may limit generalization to a more generic and diverse population [1], [9]. Multi-centre heart data or real-time hospital records would make the application more clinically reliable. Future work adding OCR methods to read in scanned and handwritten medical reports will allow for full automation. More advanced models, such as CNNs and recurrent models, work on achieving better predictive accuracy and learning feature representation. Furthermore, using explainable AI approaches such as SHAP or LIME might yield more insight into model decision making for better trust and engagement by health professionals. Integration with devices and telemedicine infrastructure is another approach to continuous cardiac health monitoring that shows great promise.

REFERENCES

[1] Detrano, R., Janosi, A., Stein Brunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International operation of a new probability algorithm for the opinion of coronary roadway complaint. The American Journal of Cardiology,64(5),304–310.
[2] Gudadhe, M., Wankhade, S., & Dongre, S. (2010). Decision support system for heart disease grounded on support vector machine and artificial neural network. In Proceedings of ICCCT (pp.741–745).
[3] Jabbar, M. A., Deeksha Tulu, B. L., & Chandra, P. (2014). Bracket of heart disease using K- nearest neighbour and inheritable algorithm. In Proceedings of ICECIT (pp. 24 – 29).
[4] Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical lores, 3(3), 290 – 294.
[5] Ayon, S. I., & Islam, M. M. (2019). Heart disease prediction using data mining ways A study on the UCI dataset. SmartCR, 9(2), 131 – 151.
[6] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system for prognosticating heart disease using machine literacy algorithms. Mobile Information Systems, 2018.
[7] Pedregosa, F., Varoquaux, G., Gram fort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit- learn Machine literacy in Python. Journal of Machine Learning Research, 12, 2825 – 2830.
[8] OpenAI. (2024). GPT- 4/ GPT- 4o Technical Report. OpenAI Publications.
[9] UCI Machine Learning Repository. Heart Disease Dataset. Available https://archive.ics.uci.edu/