SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

CARDIOAI: An Intelligent Heart Disease Prediction System Using Machine Learning

Devesh Kumar[1], Neerjal Katoch[2], Suraj Kumar[3], Rishu Kumar Singh[4], Dr. Santosh Warpe[5]

[1,2,3,4,5] Department of Computer Engineering, MIT Academy of Engineering, Alandi (D), Pune, India.

Abstract
Heart disease still ranks among the leading causes of mortality; therefore, efficient and easily available screening is crucial for preventing deaths related to cardiovascular issues. This paper describes CardioAi – a web-based platform, which analyzes the clinical features of a patient and assesses the cardiovascular risk using machine learning. Four supervised models – Logistic Regression, Random Forest, Support Vector Machine and Decision Tree are applied to build the model based on patients' medical records, while the final prediction is provided along with the risk visualization, PDF report creation and advisory tool working on the basis of rules instead of being presented as a number. The system is built as a full-stack web application using React.js for front-end development, FastAPI for back-end development and MySQL for data storage. Two different user roles are considered: Patient and Doctor, whose access is regulated via role-based authentication allowing both a patient and their physician to monitor the history of predictions. On the evaluation set consisting of more than 50,000 patient records, Random Forest classifier demonstrates the highest accuracy among four models – approximately 88% versus Support Vector Machine (~86%), Logistic Regression (~85%) and Decision Tree (~82%).

Keywords: heart disease prediction, machine learning, random forest, logistic regression, support vector machine, decision tree, risk prediction, healthcare informatics, medical advisory system.

I. Introduction

CardioAi is a machine learning based system, which takes into consideration the parameters associated with a patient's clinical profile such as the age, blood pressure, cholesterol and similar ones to find the cardiovascular risk associated with him and display the output in the form of graphical analysis along with the medical advice. The basic idea originated from a very practical issue. Health screening of a patient generally becomes difficult as he cannot reach his cardiologist in time, leading to late intervention of the disease.

In several predictable ways, current processes fail to accomplish this aim. Conventional heart disease diagnosis is almost exclusively reliant on the expertise of specialists, making it prone to human error, subjective judgment, and long waiting periods of days, as well as not being accessible to people living in remote locations without access to a cardiologist or testing facilities. Patients may also overlook signs that could indicate potential problems due to their unawareness. Although digital aids are available, most of them lack the machine learning ability to provide a prediction, maintain the patient's prediction history in a single location, and give any guidance after the risk level has been assessed.

In order to bridge these gaps, the proposed solution trains and evaluates four classic machine learning classification algorithms based on clinical patient data and feeds the prediction to a risk assessment and advisory pipeline rather than presenting the model output directly. In case the predicted risk is high, the solution raises an alarm; otherwise, it provides preventive advice. The whole pipeline has been developed as a full-stack web application with Patient and Doctor modules, user authorization, visual analytics, and PDF reports generation, with the ultimate aim of making cardiovascular screening convenient, cost-effective, and uniform compared to manual procedures.

Many of the predictive models developed at the prototype level consist only of a single classifier and the resulting printed out accuracy. CardioAi is designed and tested as an end-to-end workflow: the very same system that generates the percentage risk also generates the corresponding role-based action advice and maintains a persistent history that the physician will be able to review along with the patient. A prediction that does not lead anywhere or is never seen by the physician who treats the patient offers little over no prediction at all.

II. Literature Survey

Deep learning with electronic health records. Rajkomar et al. demonstrate that models based on deep learning using large-scale electronic health records can attain very high levels of prediction accuracy for various clinical outcomes [15]. The findings suggest that model-based predictions obtained directly from the patient data can be at least as accurate as those derived from risk scores, which is one reason why this work employs a fully learned classifier pipeline instead of a clinical scoring rule.

Feature selection and ensemble learning. Pandey and Jain incorporate feature selection into ensemble learning for enhanced accuracy of heart disease classifications [16], while Shetty and Aggarwal design a hybrid method of feature selection which enables dimensionality reduction and removal of noise from the clinical data sets [14]. These findings lead us to the conclusion that proper feature selection, coupled with ensemble classifiers, performs better than a model using raw features, which influenced CardioAi's feature selection phase before modeling.

Feature Engineering and Classical Classifiers. Feature engineering and classical classifiers, such as logistic regression, decision trees, and support vector machines, have been examined by Amin et al. and by Patel and Mehta to determine their influence on predictive performance, and both find that ensemble methods tend to outperform individual classifiers [12], [13]. This is one reason why multiple classical models were trained and compared to each other rather than only one.

Deep learning and data mining techniques. Khan et al. continue research in this area using deep neural networks to score patients' cardiovascular risks [10], while Srinivas and Raghavendra Rao use traditional data mining methods to find the best predictors for heart disease [11]. Both papers show that it is important to analyze features systematically prior to classifying, and CardioAi does it as a pre-processing step.

Ensemble learning techniques and entropy-based selection of features. Recent research has taken these directions even further. IntelliHealth is an ensemble classifier created by Absar et al., which incorporates multiple classifiers to increase the robustness of the system [8]. Entropy-based feature selection has been proposed by Rajendran to extract the most informative clinical features [9], while Bhatt et al. compare the performance of automated machine-learning pipelines to classical models in diagnosing diseases [4]. All these results confirm the selection of Random Forest classifier in CardioAi.

New hybrid, optimization and explainable AI approaches. Recent advancements continue to explore improvements in terms of both accuracy and interpretability: Menshawi et al. construct a hybrid diagnostic system based on multiple algorithms [5]; Sumwiza et al. enhance Random Forest ensembles through efficient feature selection [6]; and Kanchanamala et al. develop a hybrid optimization-based approach to constructing convolutions for heart disease prediction [7]. Bouqentar et al. and Liu et al. provide an overview of feature engineering and of machine learning cardiovascular prediction models respectively [1], [2], and El-Sofany et al. incorporate explainable AI techniques to increase transparency of predictions [3]. The medical advisory layer of CardioAi, providing interpretability of the numeric risk score in a form of understandable advice, is the result of such an approach.

Consistently, throughout the entire body of literature analyzed, there is an approach that proves superior: that is, the use of ensembles over individual classifiers, and the employment of feature selection in combination with multiple algorithms. This is precisely the strategy adopted in the suggested system.

III. Proposed Methodology

A. System Overview

This is accomplished via a series of stages such as data entry of patients, data preprocessing, feature selection, multiple models for predictions, medical advice based on risk and visualization of the reports. This architecture is shown in Fig. 1, where the entire pipeline from the data capturing stage at the front end up until the dashboard and reporting stage is depicted, integrating the Patient and Doctor portals with the Prediction Engine, Risk Scoring Engine, Medical Advisory Module and Report & Dashboard Module through a Flask/FastAPI REST backend and a MySQL data store.

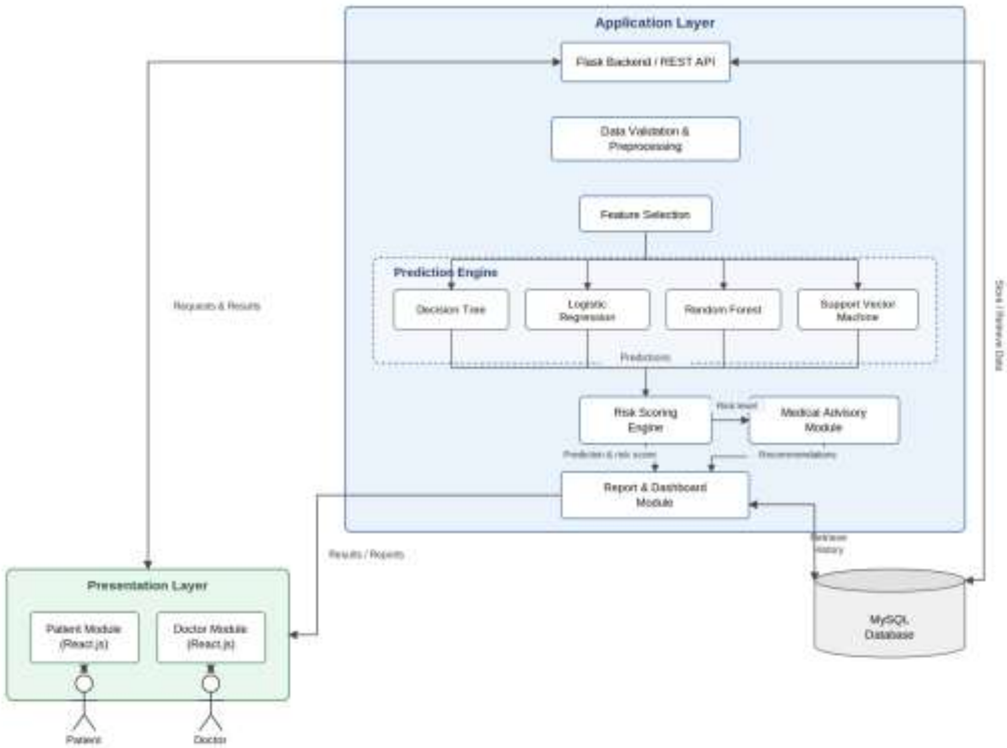

Fig. 1. System architecture of the CardioAi prediction pipeline.

B. Acquisition and Preprocessing of Data

The health data of the patients is acquired via a React.js form at the time of their enrollment or follow-up visits. For this particular research project, the models were trained and tested using a preprocessed data set consisting of roughly 50,000 records. This provided the classifiers with a sufficiently large data pool to learn from. The processed data is normalized prior to its input into the classifier to ensure that there is uniformity in the input regardless of how the raw values have been entered.

C. Feature Selection and Model Training

The important features for prediction are selected based on the feature selection process that uses the patterns explained in Section II. The four classifiers that are Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree classifiers are trained using the same feature set of 50,000 records with the help of Scikit-learn.

D. Prediction Engine and Risk Scoring

During the inference phase, the models output their prediction, which is transformed into the form of risk in percentage format. Random Forest classifier, which turned out to be the best performing one in the assessment phase (see Section IV), is used as the main scoring model, while the rest of classifiers remain accessible via the Prediction API.

E. Medical Advisory System

The risk level determined will influence the type of advice generated by the system. In cases where the patient is high-risk, the system generates warnings immediately, advises the patient to consult with a doctor urgently, provides emergency contact information, and offers guidelines on monitoring the symptoms. Where the patient is low risk, on the other hand, the system generates advice regarding preventive healthcare, scheduling of the patient's visits to a hospital, dietary and lifestyle advice, and exercise advice for maintaining a healthy heart.

F. Web Dashboard and Deployment

Predictions, percentage of risks, and historical data trends are presented using the React.js dashboard which is powered by the FastAPI backend and MySQL database. In the Patient Module, tasks such as registration of patients, filling of patient health data, prediction result viewing, and downloading of the report in PDF format are performed, whereas, in the Doctor Module, the physician has the authority to view the report of the patient, predictions, make recommendations, and view patient history.

G. Implementation Stack

The pipeline has been developed using Python and Scikit-learn to develop the model, FastAPI for the Prediction and Advisory APIs, React.js for both the Patient and Doctor interfaces, and MySQL for storing patient information, history of predictions, and reports generated. The integration of the four models, risk scoring, and advisory decision logic under one FastAPI microservice ensures that the front end remains just a presentation layer and also the underlying model can be changed easily without modifying the Patient or Doctor modules.

IV. Results

The four classifier models were trained and tested on the same preprocessed dataset of about 50,000 patient records for their comparison of accuracy. Table I gives the results of the evaluation.

TABLE I. ACCURACY COMPARISON OF EVALUATED ML ALGORITHMS (n ≈ 50,000 RECORDS)

Algorithm	Accuracy (%)
Logistic Regression	~85
Random Forest	~88
Support Vector Machine	~86
Decision Tree	~82

In terms of accuracy, Random Forest is the most accurate among all four models with approximately 88% accuracy followed by Support Vector Machine (~86%), Logistic Regression (~85%) and Decision Tree (~82%) as can be seen from Table I. The above order corresponds to the advantage of ensembles that has been discussed in Section II and is consistent with the use of Random Forest as the main score prediction model of CardioAi. Apart from the accuracy of models, the evaluation of our prototype from end-to-end includes the following: role based login system for patients and doctors, personalized dashboard, prediction based on machine learning model along with its risk percentage, visual representation of risk, health report in PDF format, and finally, rule-based medical advisory response for both high-risk and low-risk cases.

The difference between Random Forest and the other algorithms is not too much – only up to six percentage points on the ~50,000-record test dataset – meaning that, for the given set of features, the choice of the algorithm does not have as much importance as the clinical data used. Lower accuracy of the Decision Tree algorithm can be attributed to the fact that it overfits on small sets of features by creating shallow decision boundaries. The fact that Logistic Regression and SVM are quite close to each other indicates the linear separation between classes after the feature selection.

These results also illustrate the limitations of the current evaluation, since the accuracy of the classifiers does not go beyond the low 80% range while some deep-learning models or large ensembles achieve higher numbers in scientific literature.

V. Conclusion

CardioAi was described in this paper as an intelligent system to predict heart diseases using machine learning. The system analyzed Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree algorithms and passed the calculated risk score via a medical advisory and report stream, rather than just giving a prediction. Compared to a conventional manual screening process dependent on specialists, the system provides faster, centralized, and role-based access for both patients and physicians via a full-stack web platform. In a dataset of 50,000 patients, the Random Forest algorithm achieves an accuracy of approximately 88%, which is the highest among the four classifiers used and corresponds to the findings from past literature.

VI. FUTURE SCOPE

There is much scope for further research in relation to this current prototype design. Firstly, the size of the ~50,000 records database, which is bigger compared to other similar studies, can be made even bigger with data obtained from different hospitals and demographic profiles so that it would become less biased and generalize better on various patients. Secondly, deep learning approaches, e.g. neural networks or hybrid CNN methods, could be researched together with the existing classical classification algorithms to determine whether they would bridge the 90%+ accuracy gap of some ensemble models.

Third, using real-time IoT wearable data (heart rate, ECG, blood oxygen level) will allow the application to shift its focus from static screening to constant cardiovascular monitoring with alarms being triggered whenever there are changes in the readings rather than only during manual data entry. Fourth, implementing this system on scalable cloud-based platforms with mobile apps for both patients and physicians will provide easier access

for users beyond the desktop web browser format and also enable Doctor Module scaling to cover more patients.
The last step would involve implementing Explainable-AI methodologies right inside the Prediction API as opposed to considering the issue of interpretability as something different, as a result of which the Doctor Module would identify what clinical characteristics led to a particular risk score being generated, thus helping the physician make an informed decision based on the number rather than just blindly trusting it.

References

[1] Mohammed Amine Bouqentar, Oumaima Terrada, Soufiane Hamida, Shawki Saleh, Driss Lamrani, Bouchaib Cherradi, Abdelhadi Raihani,, "Early heart disease prediction using feature engineering and machine learning," Procedia Computer Science / Computers & Electrical Engineering, 2024.
[2] Tianyi Liu, Andrew Krentz , Lei Lu, Vasa Curcin "Machine learning based prediction models for cardiovascular disease: A systematic review," Medical Engineering & Physics, 2024.
[3] H. El-Sofany, B. Bouallegue, and Y. M. Abd El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," Scientific Reports, vol. 14, 23277, 2024.
[4] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, vol. 16, no. 2, p. 88, 2023.
[5] A. Menshawi, M. M. Hassan, N. Allheeib, and G. Fortino, "A hybrid generic framework for heart problem diagnosis based on a machine learning paradigm," Sensors, vol. 23, no. 3, p. 1392, 2023.
[6] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," Informatics in Medicine Unlocked, vol. 41, 101316, 2023.
[7] P. Kanchanamala, A. S. Alphonse, and P. V. Bhaskar Reddy, "Heart disease prediction using hybrid optimization enabled deep learning network with spark architecture," Biomedical Signal Processing and Control, 2023.
[8] N. Absar et al., "IntelliHealth: Hybrid classifier ensembles for heart disease prediction," Journal of Biomedical Informatics, 2022.
[9] R. Rajendran and A. Karthi, "Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers," Expert Systems with Applications, vol. 207, 117882, 2022.
[10] M. Khan et al., "Deep learning for cardiovascular risk prediction," Elsevier Biomedical Engineering, 2021.
[11] K. Srinivas and G. Raghavendra Rao, "Analysis of heart disease prediction using data mining techniques," International Journal of Computer Science and Engineering, 2021.
[12] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Tehran Heart Center Journal, 2020.
[13] R. Patel and A. Mehta, "Heart disease prediction using machine learning algorithms," IEEE Access, 2020.
[14] D. K. Shetty and D. Aggarwal, "Hybrid feature selection approach for heart disease prediction," International Journal of Computer Applications, vol. 178, no. 9, 2019.
[15] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," npj Digital Medicine, vol. 1, no. 1, p. 18, 2018.
[16] P. Pandey and S. Jain, "Heart disease prediction using ensemble learning and feature selection," Biomedical Signal Processing and Control, vol. 52, pp. 24–29, 2018.