

An Intelligent Clinical Decision Support System for Cardiovascular Disease Prediction Using EHR-Based ML and DL Techniques

V. Madhusudhana* and K. Kavitha

Department of Artificial Intelligence and Machine Learning, Mohan Babu University, Tirupati, Andhra Pradesh, India

*Corresponding author: vayyavurumadhusudhana@gmail.com

Abstract. Artificial intelligence (AI) is widely used in healthcare to predict disease. Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, so early detection is vital for lowering the cost of treatment and improving patient survival. Many classification techniques have been designed for CVD prediction, but high prediction accuracy is still difficult to achieve. The main goal of this study is to increase the accuracy of CVD prediction with advanced AI techniques. Because CVD poses a great threat to the global population and early identification of patients is important, the study examines machine learning (ML) and deep learning (DL) algorithms for predicting CVD from electronic health records (EHRs). The predictive capability of conventional AI algorithms was estimated and compared on a large EHR dataset. According to the results, the proposed D-t-SNE method performed better than the ensemble models it was compared with.

Keywords. Medical data mining; electronic health records; data-driven healthcare; deep learning; gradient boosting; CatBoost; digital health; personalized healthcare; precision medicine

I. INTRODUCTION

Heart disease is one of the major contributors to mortality, so effective strategies for addressing it are important. Conventional risk assessment methods have played an important part in the prevention and management of cardiovascular disease, but they have drawbacks that limit their efficacy [1]. One major drawback is their dependence on a few demographic and primary health factors. Their accuracy is limited when they are applied to populations with varied genetic backgrounds, environmental conditions, and health services. Conventional risk assessment also ignores the fact that risk changes over time with lifestyle, newly acquired medical conditions, and environmental conditions [2].

Conventional methods have mainly identified people at risk from the presence of risk factors, but they do not recommend how those factors could be modified or which prevention strategies suit an individual [3]. A more thorough, dynamic, and customized risk assessment that examines each person's risk profile could improve cardiovascular outcomes across populations. Traditional risk estimators follow a one-size-fits-all strategy and ignore the traits of different populations, although lifestyle plays a large role in defining cardiovascular risk [4]. A hybrid risk assessment tool should therefore address four areas. *Personalization*: risk categories are tailored to individual characteristics and lifestyle decisions, so that recommendations can be given. *Modifiable factors*: individuals can recognize the risk factors they are able to change and take control of their health

management. *Combined strengths*: several estimators are combined, so a larger number of risk factors is considered and accuracy improves [5]. *Efficiency*: the tool is designed to work efficiently in modern healthcare settings.

II. RELATED WORK

Peng and colleagues [6] introduced the XGBH method, which uses the histogram technique to capture the critical factors and improve prediction performance. Their results on the Chinese Shangui dataset and a Kaggle dataset showed a significant improvement in prediction accuracy when body mass index (BMI) was included.

Pal and colleagues [7] used a University of California sample to evaluate the multilayer perceptron (MLP) and k-nearest neighbors (KNN) techniques. After outliers and missing values were handled, the MLP gave better prediction accuracy than KNN. Rezaianzadeh and colleagues [8] examined the connection between gender-specific death rates in a southern Iranian city and modifiable cardiovascular risk factors such as high blood pressure, diabetes, and alcohol consumption, using log-rank tests and bias-reduced Cox regression analysis. Vaduganathan and colleagues [9] stressed the relevance of modifiable and non-modifiable risk factors in the development of cardiovascular disease. Blankenberg and colleagues [10] identified five major modifiable risk factors that contribute to the global burden of cardiovascular disease, and they used Cox regression models to estimate ten-year cardiovascular risk from disease

incidence and mortality. Peng and colleagues [6] also found that cholesterol level, blood pressure, and age have a large effect on cardiovascular disease prediction, and that adding BMI to the XGBH model improved model stability and prediction accuracy. Zhou and colleagues [11] proposed the P-CARDIAC model for the Chinese population and stressed the need to account for medication interactions and population characteristics.

III. RESEARCH METHODOLOGY

The workflow of the study is shown in Fig. 1. The first step is to gather large amounts of data that cover lifestyle, genetic, and environmental factors [12]. The data are then pre-processed to make them reliable. Structuring the data allows meaningful analysis and the identification of cardiovascular disease trends in different population groups.

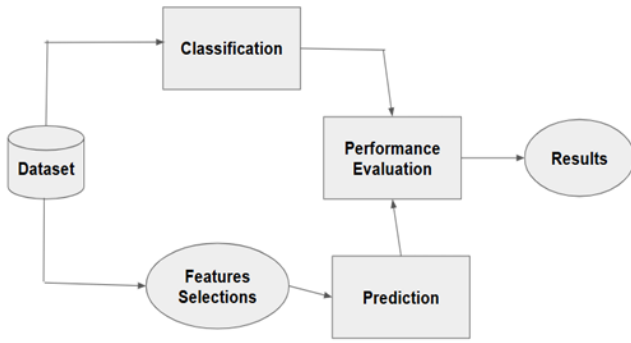


Fig. 1. Workflow diagram.

Sophisticated analytical tools are needed to analyze complex health data and make predictions. Evaluation increases the accuracy of the predictions and makes them useful for action. User feedback also improves accuracy, because it keeps the predictions up to date with current public health trends [13].

The predictions are tested to check the accuracy of the predictive tools, and intensive testing is intended to maintain accuracy regardless of the amount of data available. Adaptive evaluation methods improve the interventions that follow from the predictions, which benefits the health of individuals and groups. Effective data collection, evaluation, reporting, testing, and adaptive evaluation build trust among individuals and promote healthy behavior [14]. Collaboration among stakeholders in data collection can improve individual cardiovascular health and, in turn, public health outcomes.

A. Compared Models

The ML and DL models compared in this study compute the prediction $\hat{y}$ as follows. For LightGBM, gradient boosting, and CatBoost, $f_m(x)$ is the output of the m th of M base learners for the input x , and γ_m is its weight.

LightGBM:

$$\hat{y} = \sum_{m=1}^M f_m(x). \quad (1)$$

TabNet:

$$\hat{y} = \text{softmax}(h(X)), \quad (2)$$

$$M^{(t)} = \text{sparsemax}(P^{(t-1)} \odot h^{(t)}(a^{(t-1)})). \quad (3)$$

Gradient boosting:

$$\hat{y} = \sum_{m=1}^M \gamma_m f_m(x). \quad (4)$$

Random forest, with N trees f_i :

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N f_i(x). \quad (5)$$

CatBoost:

$$\hat{y} = \sum_{m=1}^M f_m(x). \quad (6)$$

AdaBoost, with classifier weights α_m :

$$\hat{y} = \text{sign}\left(\sum_{m=1}^M \alpha_m f_m(x)\right). \quad (7)$$

LSTM, where x_t and h_t are the input and hidden state at step t , W_h and W_x are weight matrices, and b is the bias:

$$h_t = f(W_h h_{t-1} + W_x x_t + b). \quad (8)$$

B. Proposed Work

Fig. 2 shows the overall flow of the proposed method, and Algorithm 1 gives the D-t-SNE procedure. First, the Cleveland dataset is imported, and data cleaning removes superfluous records. To increase the efficacy and reliability of the classifier, feature extraction and fusion by two deep convolutional neural network (CNN) models identify a useful set of features. The dimensionality is then reduced with the proposed D-t-SNE, which also reduces the data storage space. D-t-SNE was introduced to improve classifier performance in heart disease prediction by eliminating superfluous data.

ALGORITHM 1. D-T-SNE

Input: dataset $D = \{p_1, p_2, \dots, p_n\} \in \mathbb{R}^C$, perplexity $perp$, learning rate η , number of iterations H , and momentum $\alpha(h)$

Output: low-dimensional dataset $E^{(H)}$

- 1: Compute the high-dimensional similarities with perplexity $perp$
- 2: Build the matrix T of high-dimensional similarities
- 3: Initialize $E^{(0)}$ from $\mathcal{N}(0, 10^{-4}I)$
- 4: **for** $h = 1$ to H **do**
- 5: Compute the low-dimensional similarities

$$S_{u,v} = \frac{(1 + \|E_u - E_v\|^2)^{-1}}{\sum_{k \neq l} (1 + \|E_k - E_l\|^2)^{-1}}$$
- 6: Build the matrix S
- 7: Estimate the gradient $\partial C / \partial E$
- 8: Set $E^{(h)} = E^{(h-1)} + \eta \partial C / \partial E$

$$+ \alpha(h) (E^{(h-1)} - E^{(h-2)})$$
- 9: **end for**
- 10: Output the low-dimensional dataset $E^{(H)}$

A hyperparameter-tuned MLP (H-MLP) is then used for classification. Neural networks such as MLPs, recurrent neural networks (RNNs), and CNNs are widely used for this task. Hyperparameters govern how the model parameters are learned, so different hyperparameter values produce different model parameters for the same data. Hyperparameter

tuning identifies the best set of hyperparameter values and applies the algorithm with them, which reduces computation time and helps in visualizing the data. The H-MLP

classifies each record by the presence or absence of the disease. Finally, the trained model makes the predictions, and a performance analysis measures the efficiency of the system.

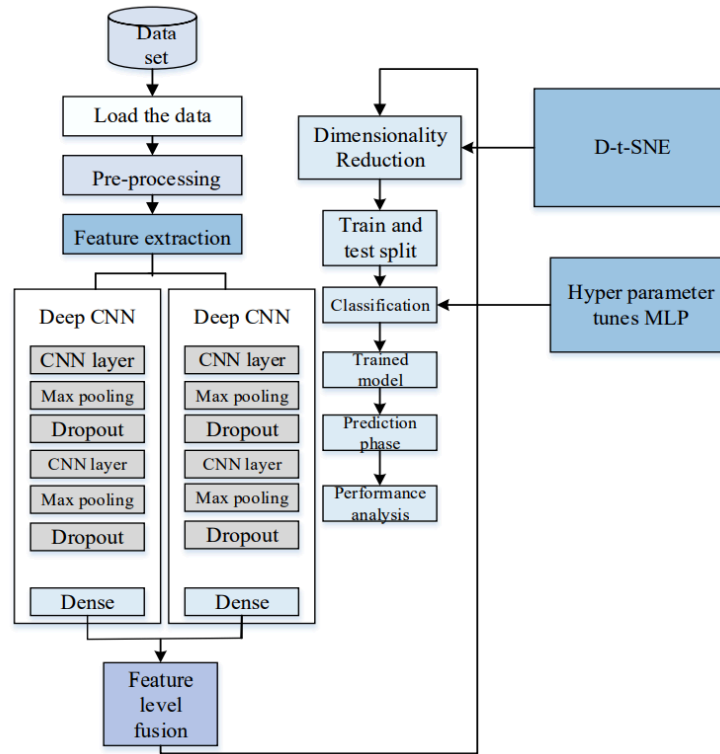


Fig. 2. Proposed method.

IV. RESULTS AND DISCUSSION

This study proposes a CVD prediction system intended to improve healthcare management and patient outcomes. Data from a heterogeneous population are collected, and varied health information forms a large database for analysis. The database includes many cardiovascular risk factors, such as demographic, lifestyle, and clinical information, which are used to detect patterns of cardiovascular risk and predict the disease.

Data preprocessing was important for detecting and categorizing the major risk indicators and for lowering the likelihood of incorrect predictions. The process also includes approaches for continuous evaluation, which make it sensitive and adaptable to new information. The proposed model was designed to keep its prediction accuracy over time when it deals with large amounts of dynamic, changing healthcare data. Tests showed the efficacy of the method under different data conditions, which is valuable in real-time healthcare applications where data quality, data availability, and patient information can change quickly.

Different ML and DL methods were applied to identify the disease, and their performance was compared (Fig. 3). LightGBM, a fast gradient boosting framework for large volumes of data, achieved 85.4% accuracy. TabNet, which uses attention to extract the significant features from tabular data, achieved 84.2%. Gradient boosting, which trains sequential weak learners on past errors, achieved 83.6%. Random forest, which combines many decision trees to improve stability and reduce overfitting, achieved 82.9%. CatBoost,

which handles categorical features with minimal preparation, achieved 86.1%. AdaBoost, which combines weak classifiers and focuses on misclassified instances, achieved 81.7%. The long short-term memory (LSTM) network, a recurrent neural network for sequence prediction, achieved 88.3%. D-t-SNE performed better than all the other models, with 96.68% accuracy, 96.79% sensitivity, 96.51% specificity, 97.27% precision, 96.89% F1 score, and 93.24% Matthews correlation coefficient (MCC).

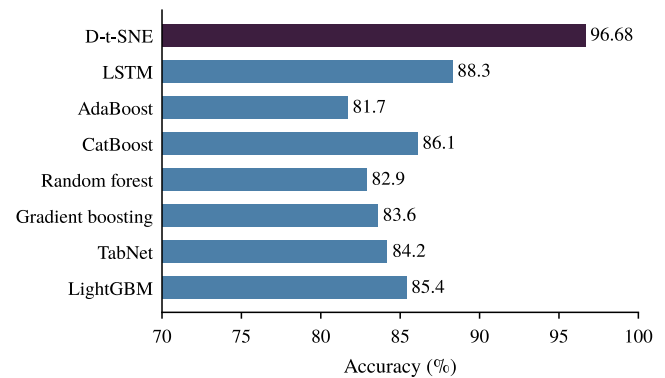


Fig. 3. Accuracy (%) of the compared models.

V. CONCLUSION

Identifying cardiovascular disease is a major problem in healthcare, and predicting heart disease accurately is difficult. This paper applied several ML techniques to the prediction of cardiac disease. Large volumes of data are hard for such models to handle, so a stacked ensemble method was proposed that improves performance without requiring

a high-dimensional dataset. Heart disease can be predicted accurately with a DL model, which can help doctors treat patients. The proposed system reached an accuracy of 96.68% and performed better than the traditional approaches, allowing early detection and intervention through varied data types and analysis. The approach shows how collaboration can help reduce CVD cases. Future work will extend the system with the Internet of Things (IoT).

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] Mohammad Iqbal, Muhammad Nanda Setiawan, Mohammad Isa Irawan, Ku Muhammad Naim Ku Khalif, Noryanti Muhammad, and Mohd Khairul Bazli Mohd Aziz, "Cardiovascular disease detection from high utility rare rule mining," *Artif. Intell. Med.*, vol. 131, Art. no. 102347, 2022, doi: 10.1016/j.artmed.2022.102347.
- [2] C. Kotteeswaran, Pamidi Ankitha, and Panjagalla Veera Harish, "Cardiovascular disease prediction from retinal images using deep learning," in *Proc. 2026 8th Int. Conf. Intell. Sustain. Syst. (ICISS)*, 2026, pp. 1-8, doi: 10.1109/ICISS67859.2026.11453563.
- [3] Amar Shukla, Ginika Mahajan, Shamik Tiwari, and Akhilesh Kumar Sharma, "Metaheuristic optimized random forest for multiclass cardiovascular disease prediction," in *Proc. 2026 IEEE Madhya Pradesh Section Conf. (MPCON)*, 2026, pp. 1610-1614, doi: 10.1109/MPCON69668.2026.11509026.
- [4] Mithila Mehedi, Nourin Nishat, Janifer Nahar, Muniroopesh Raasetti, A. S. M. Shoaib, Basharat Ali, and Md Amran Hossen Pabel, "Innovative machine learning approaches to cardiovascular disease prediction," in *Proc. 2024 Int. Conf. Intell. Comput. Sustain. Innov. Technol. (IC-SIT)*, 2024, pp. 1-4, doi: 10.1109/IC-SIT63503.2024.10862741.
- [5] C. Beulah Christalin Latha and S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Inform. Med. Unlocked*, vol. 16, Art. no. 100203, 2019, doi: 10.1016/j.imu.2019.100203.
- [6] Mengxiao Peng, Fan Hou, Zhixiang Cheng, Tongtong Shen, Kaixian Liu, Cai Zhao, and Wen Zheng, "Prediction of cardiovascular disease risk based on major contributing features," *Sci. Rep.*, vol. 13, no. 1, Art. no. 4778, 2023, doi: 10.1038/s41598-023-31870-8.
- [7] Madhumita Pal, Smita Parija, Ganapati Panda, Kuldeep Dhama, and Ranjan K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Med.*, vol. 17, no. 1, pp. 1100-1113, 2022, doi: 10.1515/med-2022-0508.
- [8] Abbas Rezaianzadeh, Leila Moftakhar, Mozghan Seif, Masoumeh Ghodousi Johari, Seyed Vahid Hosseini, and Seyed Sina Dehghani, "Incidence and risk factors of cardiovascular disease among population aged 40-70 years: A population-based cohort study in the South of Iran," *Trop. Med. Health*, vol. 51, no. 1, Art. no. 35, 2023, doi: 10.1186/s41182-023-00527-7.
- [9] Muthiah Vaduganathan, George A. Mensah, Justine Varieur Turco, Valentin Fuster, and Gregory A. Roth, "The global burden of cardiovascular diseases and risk: A compass for future health," *J. Am. Coll. Cardiol.*, vol. 80, no. 25, pp. 2361-2371, 2022, doi: 10.1016/j.jacc.2022.11.005.
- [10] S. Blankenberg, C. Magnussen, F. Ojeda, and D. Leong, "Global effect of modifiable risk factors on cardiovascular disease and mortality," *N. Engl. J. Med.*, vol. 389, no. 14, pp. 1273-1285, 2023, doi: 10.1056/NEJMoa2206916.
- [11] Yekai Zhou, Celia Jiaxi Lin, Qiuyan Yu, Joseph Edgar Blais, Eric Yuk Fai Wan, Marco Lee, Emmanuel Wong, David Chung-Wah Siu, Vincent Wong, Esther Wai Yin Chan, Tak-Wah Lam, William Chui, Ian Chi Kei Wong, Ruibang Luo, and Celine Sze Ling Chui, "Development and validation of risk prediction model for recurrent cardiovascular events among Chinese: The personalized cardiovascular disease risk assessment for Chinese model," *Eur. Heart J. Digit. Health*, vol. 5, no. 3, pp. 363-370, 2024, doi: 10.1093/ehjdh/ztae018.
- [12] Reshma S. Ravi, K. Vanitha, and Anjana T. K., "A survey of cardiovascular disease identification and prediction using machine learning and deep learning methods," in *Proc. 2024 Int. Conf. Emerg. Res. Comput. Sci. (ICERCS)*, 2024, pp. 1-5, doi: 10.1109/ICERCS63125.2024.10895641.
- [13] Siddh Garg, "Cardiovascular disease prediction using classical machine learning," in *Proc. 2024 Int. Conf. Electr. Comput. Eng. Res. (ICECER)*, 2024, pp. 1-5, doi: 10.1109/ICECER62944.2024.10920302.
- [14] Surjeet Dalal, Pallavi Goel, Edeh Michael Onyema, Adnan Alharbi, Amena Mahmoud, Majed A. Algarni, and Halifa Awal, "Application of machine learning for cardiovascular disease risk prediction," *Comput. Intell. Neurosci.*, vol. 2023, Art. no. 9418666, 2023, doi: 10.1155/2023/9418666.