IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

Study on Deep Learning and Traditional Data Mining Techniques for Heart Disease and Stroke Prediction

Manoj Kumar Yadav[1] and Dr. Nilesh Vasant Ingale[2]
[1]Research Scholar, [2]Associate Professor
Department of Computer Science
Vikrant University, Gwalior, M.P, India

Abstract: *Cardiovascular diseases, including heart disease and stroke, remain among the most important causes of mortality and long-term disability worldwide. The increasing availability of electronic health records, clinical databases, medical images, electrocardiogram signals, wearable sensor data, and lifestyle information has created significant opportunities for the application of data mining, machine learning, and deep learning techniques in early disease prediction. Traditional data mining methods such as Decision Trees, Naïve Bayes, Logistic Regression, k-Nearest Neighbour, Support Vector Machines, Random Forest, and ensemble techniques have demonstrated considerable usefulness in analysing structured clinical data.*

Keywords: Heart Disease Prediction, Stroke Prediction, Deep Learning, Data Mining, Machine Learning, Cardiovascular Disease.

I. INTRODUCTION

Cardiovascular diseases (CVDs) represent a major global public health challenge. Heart disease and stroke account for a substantial proportion of cardiovascular morbidity and mortality, making early identification of high-risk individuals an important objective of modern healthcare. According to the World Health Organization, cardiovascular diseases caused an estimated 19.8 million deaths in 2022, and approximately 85% of these deaths were associated with heart attack and stroke. Early identification of cardiovascular risk can support preventive intervention, lifestyle modification, clinical monitoring, and timely treatment.

Conventional cardiovascular risk assessment has traditionally depended on clinical examination, laboratory investigations, medical history, and statistical risk-scoring systems. Although these approaches remain important, they may not fully capture the complex interactions among demographic, behavioural, physiological, biochemical, genetic, and clinical variables. Cardiovascular disorders are influenced by multiple interacting risk factors, including age, sex, hypertension, diabetes, obesity, cholesterol levels, smoking, physical inactivity, family history, and lifestyle characteristics. These interactions are frequently nonlinear, making advanced computational approaches increasingly attractive.

Data mining and machine learning provide computational methods for identifying hidden patterns and predictive relationships in large datasets. Traditional data mining algorithms have been extensively applied to structured datasets containing patient characteristics and clinical variables. Decision Trees, Naïve Bayes, Support Vector Machines, Logistic Regression, Random Forest, and k-Nearest Neighbour algorithms have been widely investigated for cardiovascular prediction.

The development of deep learning has expanded these possibilities. Deep neural architectures can automatically learn hierarchical feature representations from complex data. Convolutional Neural Networks are useful for medical imaging and ECG data, while Recurrent Neural Networks and Long Short-Term Memory models are particularly suitable for sequential and longitudinal patient information. Recent reviews have also highlighted the increasing use of hybrid architectures, multimodal models, explainable artificial intelligence, and federated learning in cardiovascular prediction.

The purpose of this review is to examine the relative contribution of deep learning and traditional data mining approaches to heart disease and stroke prediction. Particular attention is given to algorithmic characteristics, data requirements, preprocessing procedures, performance evaluation, advantages, limitations, and future research opportunities.

II. NEED FOR HEART DISEASE AND STROKE PREDICTION

Heart disease and stroke are often associated with preventable or modifiable risk factors. Raised blood pressure, blood glucose, cholesterol, obesity, unhealthy diet, physical inactivity, and tobacco use can contribute significantly to cardiovascular risk. The early detection of individuals with elevated risk can therefore support preventive healthcare and reduce avoidable complications.

Traditional clinical diagnosis generally depends on established symptoms and measurable risk factors. However, some individuals may have underlying cardiovascular disease without obvious symptoms. Predictive systems can analyse combinations of risk variables and estimate the probability of future disease. Such systems may assist healthcare professionals by identifying high-risk patients for additional screening or intervention.

Machine learning-based prediction systems are especially relevant because healthcare data are increasing rapidly in both size and complexity. Electronic health records contain demographic information, diagnoses, medications, laboratory values, clinical notes, and longitudinal observations. Medical imaging, ECG recordings, wearable devices, and biosensors generate additional high-dimensional information. Traditional statistical approaches may have difficulty handling all these heterogeneous data sources simultaneously.

Recent evidence suggests that structured stroke prediction tasks often benefit from ensemble learning approaches, while deep learning is particularly useful for imaging-based and large-scale datasets. Nevertheless, the clinical usefulness of these systems depends not only on predictive accuracy but also on external validation, interpretability, fairness, and successful integration into healthcare practice.

III. TRADITIONAL DATA MINING TECHNIQUES FOR CARDIOVASCULAR PREDICTION

Traditional data mining methods use structured variables to discover patterns, classify patients, or estimate disease risk. These techniques remain widely used because they are computationally efficient and can perform well even when datasets are relatively small.

1. Logistic Regression

Logistic Regression is commonly used for binary classification problems such as predicting whether a patient is likely to have heart disease or experience a stroke. It estimates the relationship between independent variables and the probability of an outcome. The technique is relatively interpretable and clinically useful because the contribution of predictors can be examined through model coefficients.

However, Logistic Regression may be limited when relationships between predictors and outcomes are highly nonlinear or when complex feature interactions are important.

2. Decision Tree

Decision Tree algorithms classify observations through a sequence of decision rules. For example, a prediction model may use blood pressure, cholesterol, age, and chest pain characteristics to divide patients into different risk groups.

The major advantage of Decision Trees is interpretability. The resulting model can often be understood visually by clinicians. However, individual Decision Trees may overfit training data.

3. Naïve Bayes

Naïve Bayes is a probabilistic classification technique based on Bayes' theorem. It assumes conditional independence among predictors, an assumption that may not always hold in clinical data. Despite this limitation, Naïve Bayes is computationally efficient and may perform well on appropriately prepared datasets.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

4. k-Nearest Neighbour

The k-Nearest Neighbour algorithm classifies a new patient by examining the disease status of similar patients in the dataset. Its performance depends strongly on the selection of the value of *k*, the distance metric, and the scaling of input variables.

5. Support Vector Machine

Support Vector Machine attempts to identify an optimal boundary separating disease and non-disease classes. Kernel functions can allow the algorithm to model nonlinear relationships. SVM has been widely applied to heart disease prediction because of its effectiveness in high-dimensional classification problems.

6. Random Forest

Random Forest is an ensemble learning method that combines multiple Decision Trees. It generally improves prediction stability and reduces the risk of overfitting compared with a single tree. Random Forest can also provide estimates of feature importance.

7. Gradient Boosting and Other Ensemble Methods

Boosting approaches combine multiple weak learners to produce a stronger predictive model. Recent cardiovascular prediction studies frequently report strong performance for ensemble methods, particularly when structured clinical data are available. However, performance varies according to the dataset, preprocessing strategy, class distribution, and validation method.

Table 1. Major Traditional Data Mining Techniques for Heart Disease and Stroke Prediction

Technique	Main Principle	Major Advantages	Major Limitations	Suitable Data
Logistic Regression	Estimates probability of a binary outcome	Interpretable and simple	Limited nonlinear modelling	Structured clinical data
Decision Tree	Rule-based hierarchical classification	Easy to interpret	May overfit	Structured categorical and numerical data
Naïve Bayes	Probabilistic classification	Fast and computationally efficient	Assumes feature independence	Structured clinical data
k-Nearest Neighbour	Classifies according to nearby observations	Simple and intuitive	Sensitive to scaling and noise	Small-to-medium datasets
Support Vector Machine	Finds an optimal separating boundary	Effective in high-dimensional data	Computational complexity and lower interpretability	Structured and selected-feature data
Random Forest	Combines multiple Decision Trees	Robust and strong predictive performance	Less interpretable than a single tree	Structured clinical data
Gradient Boosting	Sequentially improves weak learners	Often high predictive accuracy	Requires parameter tuning	Large structured datasets

IV. DEEP LEARNING TECHNIQUES FOR CARDIOVASCULAR PREDICTION

Deep learning is a branch of machine learning based on multilayer neural networks. Unlike many traditional approaches, deep learning can automatically learn feature representations from raw or minimally processed data.

1. Artificial Neural Networks

Artificial Neural Networks consist of interconnected layers of artificial neurons. Input features are transformed through hidden layers to generate predictions. ANN models can capture nonlinear relationships between cardiovascular risk factors.

DOI: 10.48175/568

2. Convolutional Neural Networks

CNNs are particularly useful for analysing images and spatially structured data. In cardiovascular applications, CNNs have been used for ECG interpretation, echocardiography, cardiac imaging, and other biomedical signals. Their ability to automatically identify complex visual features makes them useful when manual feature extraction is difficult.

3. Recurrent Neural Networks

RNNs are designed to process sequential information. Patient health records often contain repeated measurements over time, making sequential models potentially useful for predicting disease progression.

4. Long Short-Term Memory Networks

LSTM networks are a specialized form of recurrent neural network designed to model long-term dependencies. They can be applied to longitudinal electronic health records, repeated laboratory results, ECG sequences, and other time-series data.

5. Transformers

Transformer architectures use attention mechanisms to model relationships among different elements of a sequence. They are increasingly being investigated for complex healthcare data, including long clinical histories and multimodal information.

6. Hybrid Deep Learning Models

Hybrid approaches combine different deep learning architectures or integrate deep learning with traditional machine learning methods. Examples include CNN-LSTM models and deep learning models combined with feature-selection or ensemble techniques. Recent reviews suggest that hybrid models are becoming increasingly important in cardiovascular prediction research.

Table 2. Major Deep Learning Techniques Used in Heart Disease and Stroke Prediction

Deep Learning Technique	Main Strength	Typical Application
Artificial Neural Network	Nonlinear relationship modelling	Clinical risk prediction
CNN	Automatic spatial feature extraction	ECG and medical imaging
RNN	Sequential data modelling	Longitudinal patient records
LSTM	Long-term temporal dependency learning	ECG and repeated clinical observations
Transformer	Attention-based relationship learning	Complex clinical and multimodal datasets
CNN-LSTM	Combines spatial and temporal learning	Biosignals and sequential imaging data
Hybrid Deep Learning	Combines complementary techniques	Multimodal cardiovascular prediction

V. COMMON DATASETS AND DATA SOURCES

The effectiveness of a predictive model depends heavily on the quality and characteristics of the dataset. Heart disease and stroke prediction studies commonly use public datasets, hospital records, population surveys, imaging repositories, and electronic health records.

Structured datasets generally include variables such as age, sex, blood pressure, cholesterol, glucose, body mass index, smoking status, diabetes, chest pain characteristics, ECG findings, and family history. Stroke datasets may additionally include hypertension, heart disease history, marital status, occupation, residence type, smoking status, and average glucose level.

Large healthcare databases provide opportunities for deep learning but also create challenges related to missing values, inconsistent coding, privacy, class imbalance, and data heterogeneity.

Recent systematic reviews have identified electronic health records as a major data source for stroke prediction, followed by medical imaging and biosignal data. These reviews also emphasize that data source selection strongly influences the most appropriate algorithm.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

VI. DATA PREPROCESSING AND FEATURE ENGINEERING

Data preprocessing is a critical stage in predictive modelling. Clinical datasets often contain missing values, duplicate observations, measurement errors, inconsistent formats, and imbalanced disease classes.

Common preprocessing steps include:

1. Removal of duplicate records.
2. Treatment of missing values through imputation or exclusion.
3. Normalization and standardization of numerical variables.
4. Encoding of categorical variables.
5. Detection and treatment of outliers.
6. Feature selection.
7. Class balancing using oversampling or undersampling.
8. Division of data into training, validation, and testing sets.

Feature selection is particularly important in traditional machine learning. Methods such as correlation analysis, recursive feature elimination, mutual information, chi-square analysis, and feature importance ranking can identify the most relevant predictors.

Deep learning models may reduce the need for manual feature engineering, particularly for raw signals and images. However, preprocessing remains essential because poor-quality data can lead to unreliable models.

VII. COMPARISON BETWEEN TRADITIONAL DATA MINING AND DEEP LEARNING

Traditional data mining and deep learning should not be viewed as mutually exclusive approaches. Their relative usefulness depends on the nature of the dataset and the clinical objective.

Traditional algorithms often perform very effectively on small-to-medium structured datasets. They are computationally efficient and frequently more interpretable. Deep learning usually requires larger datasets and greater computational resources but can learn complex hierarchical features from raw and unstructured information.

Recent reviews indicate that ensemble methods frequently achieve strong results for structured heart disease and stroke prediction, whereas deep learning is increasingly advantageous for medical imaging, biosignals, and large-scale multimodal datasets. However, reported performance cannot always be compared directly because studies differ in dataset composition, outcome definitions, preprocessing procedures, and validation strategies.

Table 3. Comparative Analysis of Traditional Data Mining and Deep Learning

Parameter	Traditional Data Mining	Deep Learning
Data requirement	Low to moderate	Usually high
Feature engineering	Often required	Can be automated
Interpretability	Generally higher	Often lower
Computational requirement	Relatively low	High
Structured data	Highly suitable	Suitable
Medical imaging	Limited	Highly suitable
ECG and biosignals	Requires manual features in many cases	Highly suitable
Training time	Usually shorter	Often longer
External validation requirement	Essential	Essential
Clinical implementation	Easier when interpretable	Challenging because of complexity
Scalability to multimodal data	Moderate	High
Best use case	Structured clinical prediction	Complex, large-scale and unstructured data

VIII. EVALUATION METRICS

Model evaluation should not depend only on accuracy. In healthcare, class imbalance can make accuracy misleading. For example, a model predicting a majority non-stroke class may achieve high accuracy while failing to identify patients who are actually at risk.

Important evaluation metrics include:

1. Accuracy

Measures the proportion of correct predictions.

2. Sensitivity or Recall

Measures the proportion of actual disease cases correctly identified. High sensitivity is important when missing a high-risk patient may have serious consequences.

3. Specificity

Measures the proportion of non-disease cases correctly identified.

4. Precision

Measures the proportion of positive predictions that are actually correct.

5. F1-Score

Combines precision and recall and is useful for imbalanced classification.

6. ROC-AUC

Measures the ability of a model to distinguish between classes across multiple thresholds.

7. Precision-Recall AUC

Can be particularly useful when positive cases are rare.

In clinical prediction research, calibration and external validation are also important. A model may have good discrimination but poor calibration if predicted probabilities do not accurately correspond to real-world risk.

Table 4. Important Evaluation Metrics for Heart Disease and Stroke Prediction

Metric	Purpose	Importance
Accuracy	Overall correct predictions	General performance
Sensitivity	Identifies true disease cases	Important for reducing missed cases
Specificity	Identifies true non-disease cases	Important for reducing false alarms
Precision	Reliability of positive predictions	Useful in clinical screening
F1-Score	Balance of precision and recall	Useful for imbalanced data
ROC-AUC	Overall discrimination ability	Widely used for classification
PR-AUC	Positive-class performance	Valuable for rare disease outcomes
Calibration	Agreement between predicted and observed risk	Important for clinical decision-making

IX. CONCLUSION

The review demonstrates that both traditional data mining and deep learning techniques have significant potential for heart disease and stroke prediction. Traditional techniques such as Logistic Regression, Decision Trees, Support Vector Machines, Random Forest, and other ensemble methods remain highly valuable, particularly for structured clinical datasets. Their advantages include lower computational requirements, simpler implementation, and greater interpretability.

Deep learning approaches provide additional capabilities for analysing complex and unstructured information. CNNs are particularly useful for medical imaging and ECG analysis, while RNNs, LSTMs, and transformers can model sequential and longitudinal healthcare data. Hybrid deep learning models may provide further advantages by combining multiple learning mechanisms.

The available evidence suggests that no single algorithm is universally superior. Model selection should depend on the size, quality, type, and complexity of available data as well as the clinical objective. Ensemble methods are often highly

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, June 2026
Impact Factor: 8.2

competitive for structured prediction tasks, while deep learning becomes increasingly valuable when large-scale imaging, biosignal, or multimodal data are available. Recent reviews nevertheless emphasize persistent problems involving external validation, dataset diversity, model transparency, and clinical translation.

Therefore, future cardiovascular prediction research should prioritize hybrid and multimodal approaches, representative datasets, explainability, fairness, privacy protection, and real-world clinical validation. The most effective future systems are likely to combine the interpretability and efficiency of traditional data mining with the feature-learning capacity of deep learning. Such integrated approaches may strengthen early detection and support more personalized prevention of heart disease and stroke.

REFERENCES

1. Saba, L., & Suri, J. S. (2023). Cardiovascular disease/stroke risk stratification in deep learning framework: A review. *Cardiovascular Diagnosis and Therapy, 13*(3), 557–598.
2. Lee, F., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology, 64*(5), 304–310.
3. Dua, D., & Graff, C. (2019). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences.
4. Fathima, A. J., & Fasla, M. M. N. (2024). A comprehensive review on heart disease prognostication using different artificial intelligence algorithms. *Computer Methods in Biomechanics and Biomedical Engineering*.
5. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
6. Ghosh, S., & Reilly, D. L. (2013). Credit risk modeling using machine learning. *Proceedings of the 11th International Conference on Machine Learning and Applications*, 1–6.
7. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, 160035.
8. Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2020). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology, 69*(21), 2657–2664.
9. Kumar, P., Garg, S., Kaur, P., Johar, M., Singh, S., Menon, V., Kumar, R., Hadi, H., Hasson, R., & Lozanović, Š. (2025). A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions. *Journal of King Saud University - Computer and Information Sciences*.
10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
12. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7*, 81542–81554.
13. Olawade, D. B., et al. (2025). Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps. *International Journal of Medical Informatics, 203*, 106041.
14. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.
15. Shinde, P., Sanghavi, M., & Tran, T. A. (2025). A survey on machine learning techniques for heart disease prediction. *SN Computer Science, 6*, 334.
16. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. *IEEE Journal of Biomedical and Health Informatics, 22*(5), 1589–1604.

DOI: 10.48175/568

17. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813.
18. World Health Organization. (2025). *Cardiovascular diseases (CVDs)*. World Health Organization.
19. Yao, R., et al. (2025). Machine learning approaches for cardiovascular disease prediction: A review. *Annals of Cardiovascular Diseases*.
20. Zhang, Y., et al. (2025). A systematic review of machine learning in heart disease prediction. *[Systematic review of machine learning applications in heart disease prediction]*.