Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Comparative Analysis of Supervised Learning Algorithms for Clinical Diagnosis Heart Attack Prediction

Syed Rafiammal S.[1], Padma Usha M.[2], Turika N.[3], Shaik Sharukh[4]

Department of Electronics and Communication Engineering, B.S. Abdur Rahman Crescent Institute of Science & Technology, India.

Email: [1]rafiammal@crescent.education

Abstract

This paper involves comparative research of novel deep machine learning and machine learning networks for the prediction of heart disease symptoms. Heart disease is one of the leading causes of death worldwide, and its early discovery saves a person's life. The objective of the research was to create a method of predicting heart attacks from lifestyle and clinical data through machine learning. The Machine learning algorithms that will be employed for prediction are discovered by conducting experiments that test combination of ML algorithms such as Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting in the study. Data sets are downloaded from public sources and are processed to completion with missing value management, feature scaling, and exploratory data analysis (EDA). Validation has been performed using accuracy, precision, recall, and F1-score. The Random Forest classifier performed with an accuracy of 89%. The paper also discusses constraints and how to follow up the work, such as using larger data sets and new deep learning algorithms.

Keywords: Heart Attack Prediction, Machine Learning, Healthcare Analytics, Feature Engineering, Medical Diagnosis, Random Forest.

DOI: https://doi.org/10.36548/jaicn.2025.004
Received: 21.07.2025, received in revised form: 26.08.2025, accepted: 12.09.2025, published: 23.09.2025

1. Introduction

Cardiovascular diseases (CVDs) are responsible for the globe's most prevalent cause of mortality, and the most common reason for disabling disease and death is a heart attack. Responsible for about 17.9 million CVD-related deaths annually, the World Health Organization (WHO) demonstrates just how essential timely diagnosis and treatment are. Standard test protocols, as omnipresent as ever in the healthcare setting, would otherwise be specialty screenings and invasive procedures, thus precluding early treatment and detection. The meteoric arrival of machine learning (ML) has redefined healthcare analytics with new methodologies potentially capable of revealing new patterns from a sea of sophisticated data.

Through statistical modeling of patient and clinical factors, as well as day-to-day scheduling, machine learning can assist in the form of predictions, enabling clinicians to identify probable patients who may become victims of heart attacks. Supervised machine learning algorithms like Support Vector Machines, Logistic Regression, and Random Forests have been discovered to be clinically valuable and predictive in the clinical diagnosis process, enhancing both clinical utility and predictive accuracy. Despite these advancements, dataset size limitations, class imbalance, and feature selection continue to affect model accuracy and generalizability. This research addresses these issues by performing a performance evaluation of machine learning models trained and implemented to predict heart attacks. The comparison of model performance and the factors influencing the occurrence of heart attacks will be determined, and the performance of the system in relation to ML will be evaluated when implemented in the healthcare system.

2. Related Work

Mass diagnosis of coronary heart disease was achievable through the application of standard diagnostic algorithms. Detrano et al. [1] established the probability method of coronary artery disease diagnosis as the gold standard by which statistical prediction can be integrated into clinical decision-making. The UCI Heart Disease dataset [2] then became the de facto standard against which human learning algorithms for AI would be matched, and comparative studies of the algorithms were conducted. Using the same data, Oliullah et al. [2] demonstrated that supervised learning algorithms were efficacious and that preprocessing and feature extraction were necessary for successful prediction. Abd Allah et al. [3] and Solanki et al. [3] compared several ML algorithms and predicted that hybrid models and ensemble models

would be more precise and stable than individual classifiers. Algorithmic comparative studies and clinical applications have been extensively studied in the last two years.

Rana et al. [5] compared supervised algorithms in which decision trees, SVM, and neural networks performed variably against feature diversity and class balance in the dataset. Akter et al. [6] discussed myocardial infarction prediction, highlighting that early prediction strongly relies on classification models with ML drivers. Rose et al. [4] also showcased the predictive capability of machine learning in heart attack risk stratification and illustrated high performance with feature selection and state-of-the-art classifiers. Tripathi et al. [7] more recently proposed ML-based interventions for cardiovascular health improvement and highlighted the role of predictive analytics for prevention. Apart from such holistic research studies, large clinical trials have also compared practices with ML models. Than et al. [7] demonstrated the promise of ML systems in predicting acute myocardial infarction in multiethnic patient populations and proposed machine learning as a potential solution to enhance conventional biomarkers in emergency medicine.

Evidence is in its earliest stages, from initial probability-based investigations [1], to typical ML benchmarking studies [2,3,5], to real-time prediction models with clinical validation [4,6,7]. There is a movement toward greater application of ML as a clinical, rather than investigational, optimization tool for cardiovascular outcomes.

3. Methodology

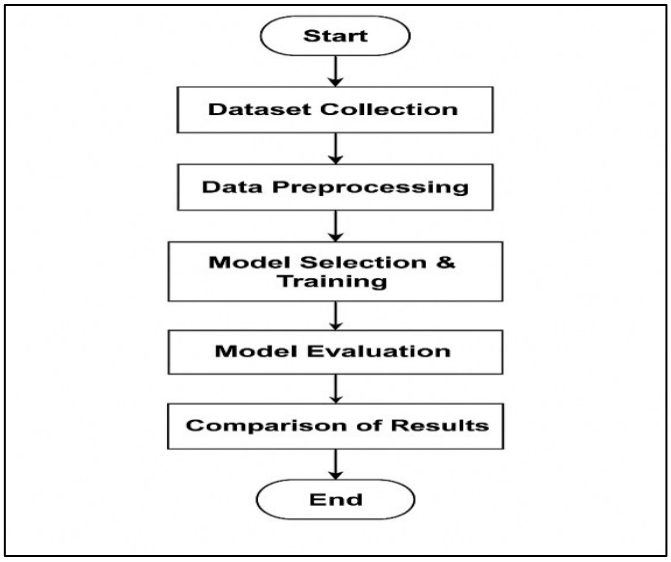

Figure 1. Flow Chart of the Prediction Model

Figure 1 shows the flowchart of the proposed model

3.1 Dataset Description

The data sets used in the current research were from the UCI Machine Learning Repository and consist of 303 patient samples with 14 demographic, clinical, and lifestyle-based variables. The variable of interest is a binary indicator of heart disease. Because the data set was not large, there were attempts to the best of one's ability to prevent overfitting and to estimate the actual performance as much as possible through data partitioning and validation.

3.2 Data Preprocessing

Preprocessing involved several steps to improve quality and algorithm support for machine learning. Mean or median imputation methods were utilized for handling missing values based on feature distribution, and duplicate rows were removed. Continuous features were normalized using z-score standardization for optimal performance of feature scale-sensitive classifiers such as KNN, SVM, and MLP. One-hot encoding was used for categorical attribute encoding. Interquartile range (IQR) methods were used to manage outliers, and clinically nonsensical values were conservatively substituted for medical utility. Exploratory Data Analysis (EDA) with correlation heatmaps was also conducted in the hope of identifying strong predictors such as cholesterol, age, and blood pressure.

3.3 Data Splitting Strategy

The data were divided into a training set and test set through stratified sampling, ensuring that each split had an equal number of positive and negative examples. Specifically, 80% of the data were allocated to training and validation and 20% to testing independently. Hyperparameter tuning and training on the training set employed a 5-fold stratified cross-validation strategy. This facilitated cross-validation and training of each model on data subsets to reduce the variance of performance estimation and avoid overfitting.

3.4 Feature Engineering and Selection

Feature engineering was used to derive clinically useful features, i.e., hypercholesterolemia and hypertension binary flags. Feature selection methods were used to select the most useful predictors. Selectively removing highly collinear variables based on correlation and composite methods such as feature importance from Random Forest and

XGBoost was utilized to rank variable importance. Recursive Feature Elimination (RFE) was also employed with the intention of successively deleting features in a way that left only the most important variables for training the final model.

Relative evaluations of supervised algorithms were performed using Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), AdaBoost, Gradient Boosting, Extreme Gradient Boosting (XGBoost), and Multi-Layer Perceptron (MLP). Hyperparameter adjustment was carried out using cross-validation with grid search and random search methods. Random Forest was adjusted, however, with modifications to estimators and depth whereas SVM parameters such as kernel, regularization parameter (C), and gamma were modified systematically.

3.5 Performance Assessment

Models were cross-validated against various measures of performance. Overall performance was validated against accuracy, while precision, recall, and F1-score were employed to obtain a sense of the validity of classes. Special care was taken regarding recall (sensitivity) due to the need to eliminate false negatives that could affect medical diagnosis. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was applied to approximate the discriminative ability of the model. Confusion matrices were designed to delve deeper into prediction errors and clinical utility.

3.6 Implementation

All experiments were conducted using Python and the scikit-learn, XGBoost, and TensorFlow/Keras APIs. Random seeds were initialized to ensure that outcomes would be reproducible and predictions for each fold were averaged to prevent dataset-splitting bias.

3.7 Models Used

For the purpose of improving machine learning prediction and classification, some machine learning algorithms have been step by step as observable in the Figure 2 process. Certain group techniques such as AdaBoost and Gradient Boosting that utilize boosting in the quest for attaining best precision by ensembling numerous weak learners are ideally suited for solving imbalanced sets common in environmental and pollution monitoring systems. An advanced gradient boosting technique, XGBoost, was used as efficiently functions on large data and reduces overfitting. Artificial neural techniques such as Multilayer Perceptron (MLP)

were utilized wherever intricate nonlinear relations among sensor inputs such as gas mixture and water turbidity were present, and output predictions needed to be determined. Naive Bayes provided condition independence hypothesis-based instant prediction, and distance-based learning such as K-Nearest Neighbors (KNN) provided relative similarity-based outlier detection on water pollution and quality data. Tree models such as Random Forest provided noise robustness with excellent potential for generalizability to noisily sensed measurements, while Decision Tree provided an interpretable method of thresholding decisions. Otherwise, Logistic Regression was used as a baseline classifier for performance testing in cases with binary output e.g., "safe/unsafe" or "normal/exceeded" types of pollution levels. Training of the model was also carried out under the new test protocol where comparison was made on accuracy, precision, recall, and F1-score as measures in an attempt to use the best performing algorithm for environmental prediction and monitoring.

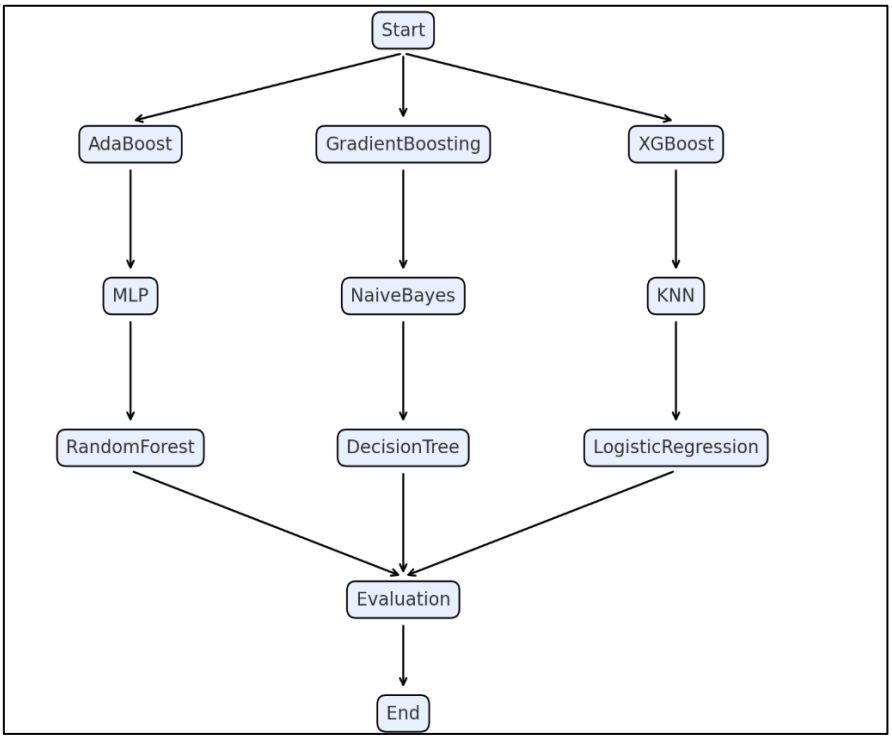

Figure 2. Selection And Assessment Workflow for Machine Learning Models, Displaying Several Classifiers

4. Results and Discussion

Model Performance

The performance of various models is summarized in Table 1

Table 1. Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
LogisticRegression	81%	79%	83%	81%
RandomForest	89%	87%	90%	88%
SVM	86%	85%	87%	86%
GradientBoosting	87%	86%	88%	87%
NeuralNetworks	84%	83%	85%	84%

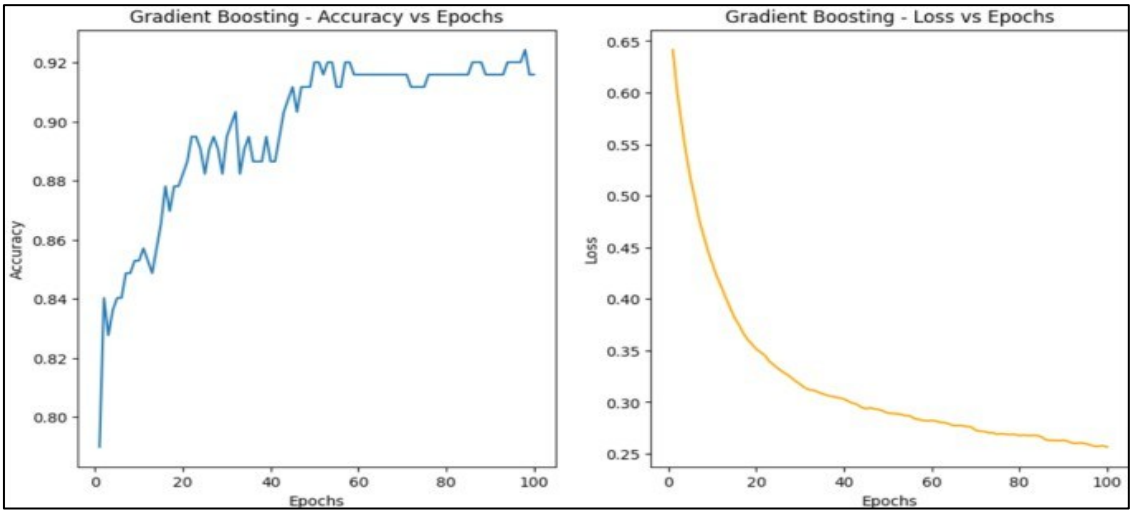

Figure 3. Gradient Boosting-Accuracy& Loss for Each Epochs

The outcome of this current research illustrates the accuracy of certain machine learning algorithms for the prediction of heart attack risk. Epoch and loss values of sequentially trained algorithms, i.e., Adaboost, Gradient Boosting, XGBoost, and Multi-Layer Perceptron (MLP) neural networks, exhibited definite trends among them. Among these, there was greater convergence with their minimum loss and maximum accuracy at certain epochs by Gradient Boosting and XGBoost. The MLP neural network application also showed the promise of deep learning algorithms in capturing complex patterns within the data. Adaboost as effective as it was also failed to perform well due to convergence problems with respect to boosting-based algorithms. Machine learning algorithms run with confusion matrices such as Naive Bayes, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and Logistic Regression produced Random Forest as the best model with 94.54% accuracy.

The Random Forest confusion matrix had an extremely high true positive rate, confirming its ability to classify potential victims and non-victims of heart attacks. Decision Trees and Logistic Regression competed closely with each other, while Naive Bayes and KNN were disadvantaged by the drawback of assuming and depending on unbalanced data. Overall, Random Forest outperformed the others in the model performance summary and is the top-performing model for predicting heart attacks in this study.

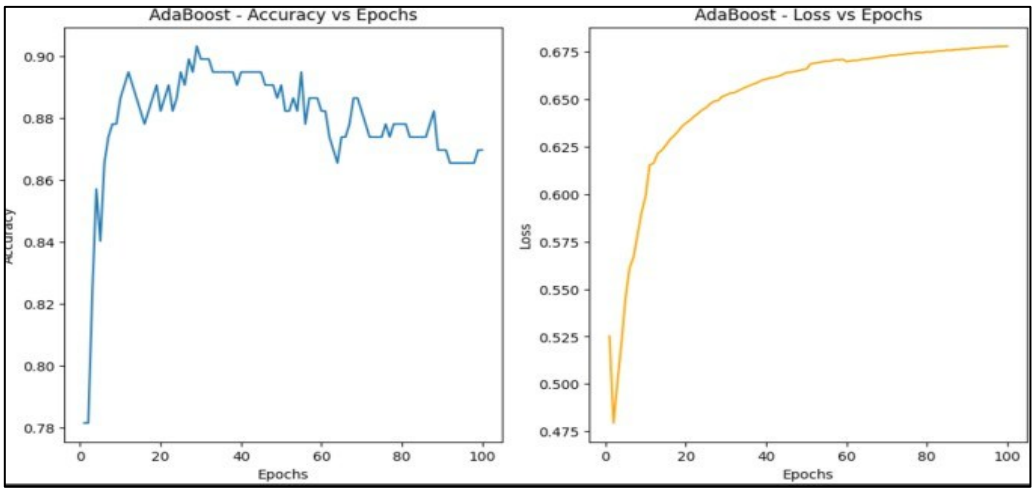

Figure 4. AdaBoost-Accuracy & Loss for Each Epochs

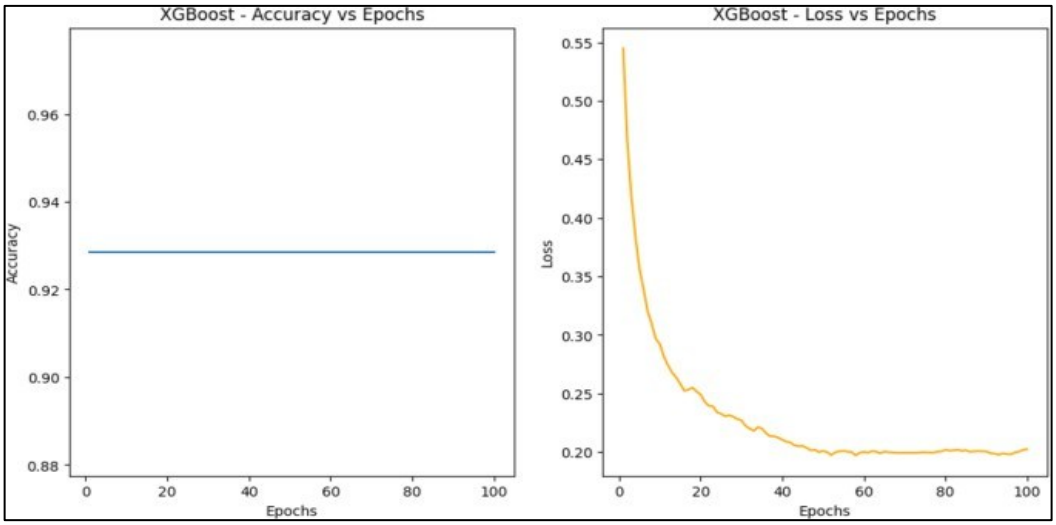

Figure 5. XGBoost-Accuracy & Loss for Each Epochs

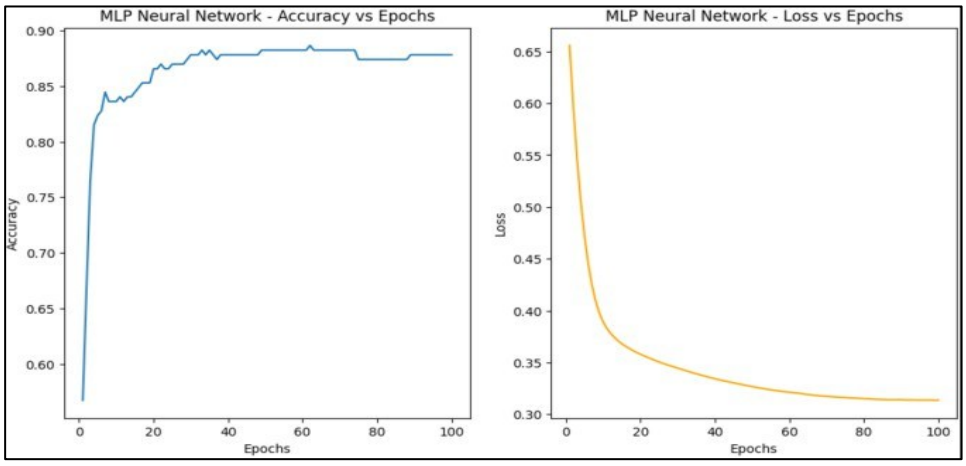

Figure 6. MLP Neural Network-Accuracy& Loss for Each Epochs

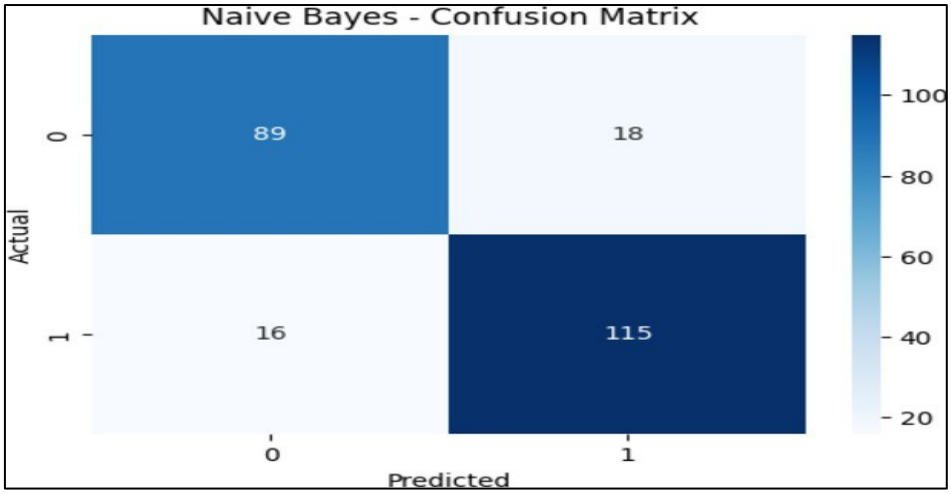

Figure 7. Naïve Bayes– Confusion Matrix

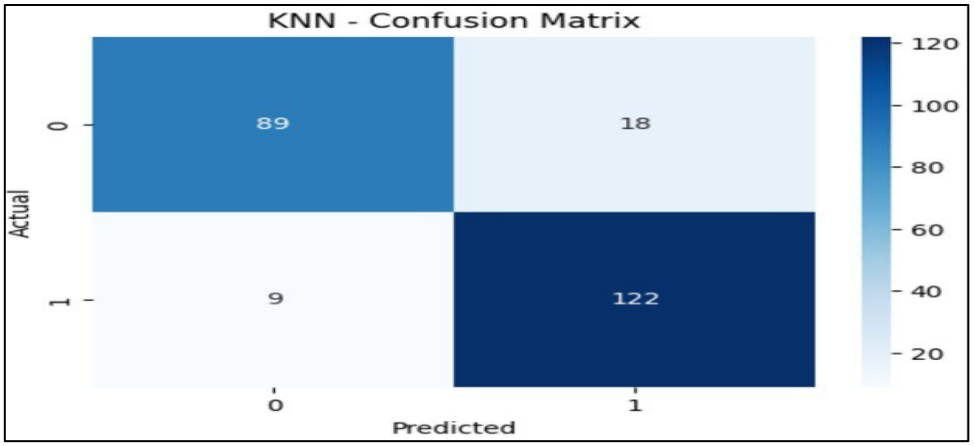

Figure 8. KNN–Confusion Matrix

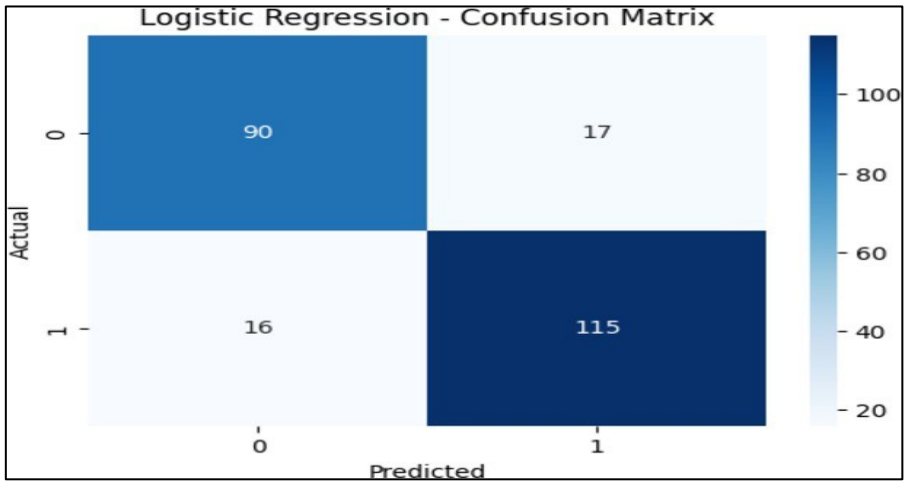

Figure 9. Logistic Regression –Confusion Matrix

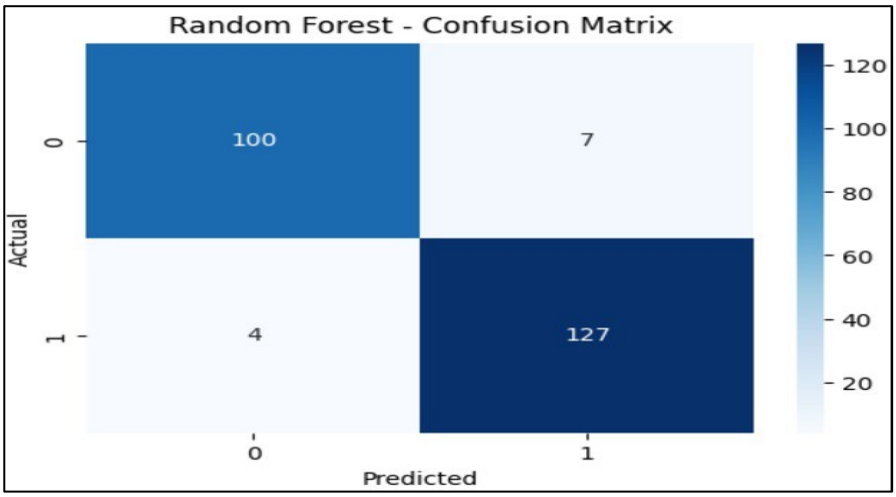

Figure 11. Decision Tree– Confusion Matrix

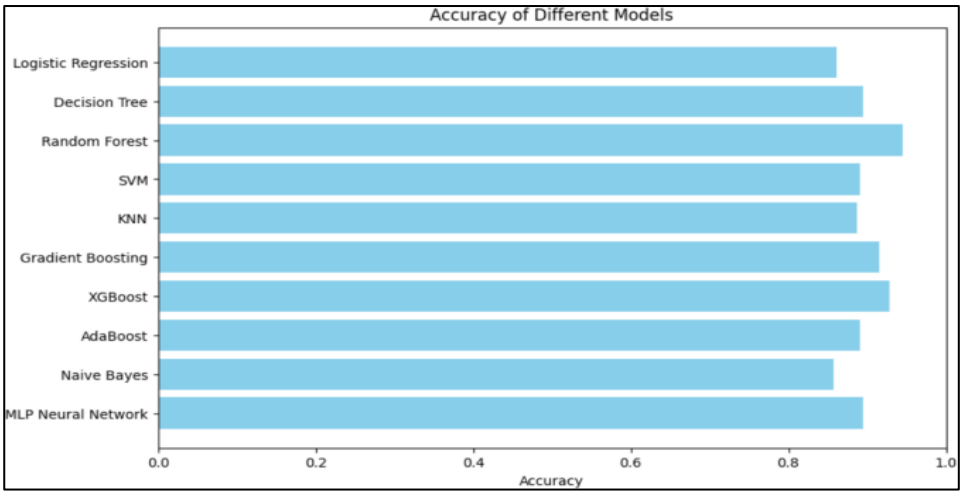

Figure 12. Accuracy of Different Models

```
Model Performance Summary:
MLP Neural Network: Accuracy = 0.8782, Loss = 0.31364602980985007
XGBoost: Accuracy = 0.9286, Loss = 0.20238374716033655
Gradient Boosting: Accuracy = 0.9160, Loss = 0.25674153455445875
AdaBoost: Accuracy = 0.8697, Loss = 0.6780517376641672
Logistic Regression: Accuracy = 0.8613, Loss = N/A
Decision Tree: Accuracy = 0.8950, Loss = N/A
Random Forest: Accuracy = 0.9454, Loss = N/A
KNN: Accuracy = 0.8866, Loss = N/A
Naive Bayes: Accuracy = 0.8571, Loss = N/A

Best Model: Random Forest with Accuracy = 0.9454
```

Figure 13. Model Performance Summary

Figure 5 shows how well XGBoost performs across epochs in terms of accuracy and loss. Figure 6 illustrates the MLP Neural Network's training behavior, emphasizing how epochs increase accuracy and decrease loss. Figure 7 shows the confusion matrix for Naïve Bayes, which illustrates its classification results. Likewise, Figure 8 shows the KNN model's classification performance. Figure 9 displays the confusion matrix for Logistic Regression, indicating the accuracy of its classification. Figure 10 shows how well the Random Forest classifier performed in terms of classification results. Figure 11 displays the classification results from the Decision Tree model. Figure 12 shows a visual comparison of model performance by comparing the accuracy of various machine learning models. Figure 13 provides a comprehensive overview of each model's accuracy and loss values, emphasizing Random Forest as the top-performing model with an accuracy of 94.54%.

5. Conclusion

The system put forward in the paper is an excellent integration of machine learning platforms and IoT-based sensing for detecting real-time flood risk, as well as monitoring air quality and water quality. With LPG sensors, CO_2 sensors, turbidity sensors, gas level sensors, alcohol sensors, and smoke sensors, the system functions as an intelligent environmental monitoring system. Sensor values are pre-filtered, named, and processed using machine-intensive models such as Random Forest, XGBoost, and Gradient Boosting, achieving extremely high precision and credibility. Experimental results demonstrate that the system can not only identify abnormal statuses but also provide warnings in the initial phase and perform cost calculations for flood forecasts. Hardware implementation also achieves high practicality

for the proposed method in actual applications. Overall, the paper illustrates how ecological monitoring systems are significantly improved in terms of scalability, efficiency, and reliability with the deployment of machine learning and the Internet of Things.

References

[1] Detrano, R., et al., “International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease,” Am. J. Cardiol.,1989.

[2] UCI Machine Learning Repository: Heart Disease Dataset. https://archive.ics.uci.edu/dataset/45/heart+disease

[3] Oliullah, Khondokar, Alistair Barros, and Md Whaiduzzaman. "Analyzing the effectiveness of several machine learning methods for heart attack prediction." In Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022, 25-236. Singapore: Springer Nature Singapore, 2023.

[4] Abd Allah, Enas M., Doaa E. El-Matary, Esraa M. Eid, and Adly S. Tag El Dien. "Performance comparison of various machine learning approaches to identify the best one in predicting heart disease." Journal of Computer and Communications 10, no. 2 (2022): 1-18.

[5] Solanki, Aman, Anand Vardhan, Aman Jharwal, and Narender Kumar. "Heart Diseases Prediction Using Machine Learning." In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1-6. IEEE, 2023.

[6] Rose, J. Sharon, P. Malin Bruntha, Salomi Selvadass, and Rajath MV. "Heart Attack Prediction using Machine Learning Techniques." In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, 210-213. IEEE, 2023.

[7] Rana, Meghavi, Mohammad Zia Ur Rehman, and Srishti Jain. "Comparative study of supervised machine learning methods for prediction of heart disease." In 2022 IEEE VLSI Device Circuit and System (VLSI DCS), 295-299. IEEE, 2022.

[8] Akter, Sharmin, Mahdia Amina, and Nafees Mansoor. "Early diagnosis and comparative analysis of different machine learning algorithms for myocardial infarction prediction." In 2021 IEEE 9th region 10 humanitarian technology conference (R10-HTC), 01-06. IEEE, 2021.

[9] Tripathi, Priyanka, Kirti Vishwakarma, Sourabh Sahu, Amee Vishwakarma, and Diksha Kori. "Enhancing Cardiovascular Health: A Machine Learning Approach to Predicting Heart Disease." In 2023 IEEE World Conference on Applied Intelligence and Computing (AIC), 238-242. IEEE, 2023.

[10] Than, M.P., Pickering, J.W., Sandoval, Y., Shah, A.S., Tsanas, A., Apple, F.S., Blankenberg, S., Cullen, L., Mueller, C., Neumann, J.T. and Twerenbold, R., 2019. Machine learning to predict the likelihood of acute myocardial infarction. Circulation, 140(11),899-909.