WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predictive Analytics of Health and Disease Detection using Machine Learning Techniques

Dr. Mrutyunjaya M S[1], Lekhana S[2], Monika A[2], Manjula K[2]

[1] Associate Professor and Head, Department of Computer Science and Engineering (Data Science) R L Jalappa Institute of Technology, Doddaballapur-561203, India.

[2] UG Student, Department of Computer Science and Engineering, R L Jalappa Institute of Technology, Doddaballapur-561203, India

ABSTRACT

Cardiovascular diseases (CVDs) are the leading cause of adult mortality worldwide, responsible for 17.9 million deaths annually, according to the World Health Organization. Machine Learning (ML) enables the extraction of valuable insights from data and is increasingly applied in health monitoring and disease prediction. This study aims to predict the likelihood of heart disease using patient medical histories, helping identify symptoms like high blood pressure and chest pain to reduce unnecessary testing. Traditional ML models such as Decision Trees, K-Nearest Neighbors (K-NN), and Naïve Bayes achieved 100% accuracy, while Support Vector Machines (SVM) showed moderate performance. XGBoost, an advanced model, reduces misclassification errors and handles data complexity effectively. Its performance can be further enhanced through feature engineering and hyperparameter tuning. Future improvements include using deep learning, ensemble techniques, explainable AI, and integrating real-time data from wearable devices, alongside privacy-preserving federated learning and cloud-based deployment for scalable, personalized healthcare solutions.

Keywords: Machine Learning, Disease Prediction, Health Monitoring, Ensemble Learning, Healthcare Analytics.

1.Introduction

The process of manipulating and extracting implicit, known or previously unknown, and possibly relevant information from data is known as machine learning. The use and scope of machine learning are always supervised learning, and unsupervised learning classifiers that are used to forecast and assess the correctness of given datasets. This information can be used to assist a huge population in programs like HDPS. Cardiovascular diseases are a broad category of heart-related ailments that are common in modern culture[1].

Cardiovascular Diseases (CVDs) account for 17.9 million deaths worldwide, according to the World Health Organization, making it the top cause of adult mortality. Our goal is to use a patient's medical history to forecast who is most likely to receive a heart disease diagnosis. It helps diagnose diseases with fewer medical tests and more effective treatments by recognizing symptoms like high blood pressure or chest pain, which contributes to prompt and focused care[2].

Four main data mining techniques are the subject of this project: Random Forest Classifier, KNN, Logistic Regression and ANN. The project outperforms earlier systems that only use one data mining technique, with an accuracy of 87.5%. This project includes the supervised learning technique of logistic regression, which works with discrete values[3].The goal is to determine a patient's likelihood of receiving a diagnosis of cardiovascular heart disease based on characteristics like age, gender, chest discomfort, fasting blood sugar, etc. The project determines if a patient may have cardiac illness by using a dataset from the UCI repository that includes patient

medical history and features. Four algorithms Random Forest Classifier, KNN, ANN, and Logistic Regression are used to train the 14 medical characteristics. Random Forest is the most effective of these, obtaining recall of 94%. Lastly, a cost-effective technique is demonstrated for classifying people at risk of heart disease[4].

2.Literature Survey

Coronary Artery Disease (CAD) is a leading cause of morbidity and mortality worldwide. It refers to the narrowing or blockage of the coronary arteries, usually due to atherosclerosis, which impairs blood flow to the heart.Traditional diagnostic methods, such as angiography, electrocardiograms, and stress testing, often have limitations, including invasiveness, cost, and reliance on clinical interpretation. Machine Learning in Medical Diagnostics: Machine learning(ML) has been increasingly applied in medical diagnostics due to its ability to process large datasets and identify complex patterns that might not be detectable through traditional methods. In the context of CAD, ML models can analyze patient data, including medical histories, biomarkers, imaging

results, and genetic information, to predict disease presence, severity, or risk factors.ML techniques, such as support vector machines (SVM), random forests, neural networks, and deep learning, have been used to improve diagnostic accuracy, enhance predictive models, and personalize treatment plans[1].

Skin diseases encompass a wide range of conditions, from mild irritations like rashes to serious conditions like melanoma or psoriasis. They can affect a person's quality of life and, in some cases, be life-threatening. Early detection and diagnosis of skin diseases, particularly malignant ones like skin cancer (melanoma), is crucial for improving treatment outcomes. However, diagnosing skin diseases accurately requires expertise and experience, making it a challenging task for general practitioners and dermatologists[2].

The dataset is diverse, sourced from multiple locations, and includes images of various skin types, lighting conditions, and camera devices, making it a robust resource for training machine learning models.The dataset is annotated with diagnostic labels by dermatologists, ensuring that each image is labeled with the correct diagnosis, which is crucial for training and validating predictive models[3].

Heart disease encompasses a variety of conditions affecting the heart, such as coronary artery disease, heart failure, arrhythmias, and valvular disorders. Early detection and diagnosis are crucial for preventing severe complications like heart attacks, strokes, and sudden cardiac arrest.Traditionally, heart disease is diagnosed through methods such as electrocardiograms(ECGs),echocardiograms, and stress tests[4].

The proposed study presents a prediction method for classifying heart disease. It distinguishes between controllable and uncontrollable risk factors. The prediction is carried out using the Random Forest machine learning algorithm[5].

The primary layer of protection for vital organs in the human body is the skin. It functions as a barrier to protect our internal organs from different sources. However, infections caused by fungus, viruses, or even dust can damage the skin. A tiny lesion on the skin can grow into something that can cause serious health problems. A good diagnosis can help the person suffering from a skin disease to recover quickly. This research aims to develop a system for detecting skin diseases using a Convolution Neural Network (CNN)[6].

It evaluates the performance of the AG-based attention U-Net. To assess the effectiveness of the framework, several state-of-the-art segmentation models are also tested on the acquired ISIC dataset[7].

An automatic algorithm designed to segment skin layers—specifically the epidermis and SLEB—in HFUS images of patients with inflammatory skin conditions such as atopic dermatitis (AD) and psoriasis. The proposed method integrates a similarity-based clustering algorithm (FCM) with a fully convolutional network and a robust post-processing step [8].

To predict the risk of developing Ischemic Heart Disease (IHD), commonly known as a heart attack, using a smartphone. An Android-based prototype application was created by incorporating clinical data collected from patients diagnosed with IHD [9].The quality parameters measured using conventional destructive methods during the storage of fresh-cut rocket leaves effectively distinguished between different visual quality levels (QL). Among these parameters, total chlorophyll and ammonia content proved to be particularly useful objective markers for evaluating all assessed QLs.[11]The performance evaluation results demonstrated that the proposed model effectively classified the different rice varieties.[20]

The primary objective of this project is to enable early and accurate rainfall forecasting, which is particularly beneficial for people living in areas prone to natural disasters like floods. It also supports farmers in making informed decisions about crop and water management. By leveraging big data analytics, this approach aims to enhance both productivity and profitability in agriculture. In this project, we propose an advanced automated framework known as the Enhanced Multiple Linear Regression Model (EMLRM), which integrates the MapReduce algorithm with the Hadoop File System.[16]

3.Methodology

3.1 Data Acquisition:

The process of gathering data from various sources for analysis. Sources can include databases, web scraping, surveys, APIs, IoT devices, or manual collection. The goal is to compile all relevant data in one place, ensuring its quality and completeness.

3.2 Data Cleaning:

The process of identifying and correcting errors, inconsistencies, or inaccuracies in the dataset. Common cleaning steps Handling missing data: Filling missing values (e.g., with mean, median) or removing incomplete records. Correcting errors: Fixing typos, outliers, or inaccuracies in data entries. Standardizing formats: Ensuring consistency in formats (e.g., date formats, capitalization). Removing duplicates: Eliminating redundant rows to avoid bias in analysis[5].

3.3 Original Dataset:

The original dataset refers to the raw, unprocessed data as it is initially collected or received. Characteristics: Contains all the data points collected from the source. Often includes irrelevant, redundant, or inconsistent information. May have errors such as missing values, noise, or outliers.

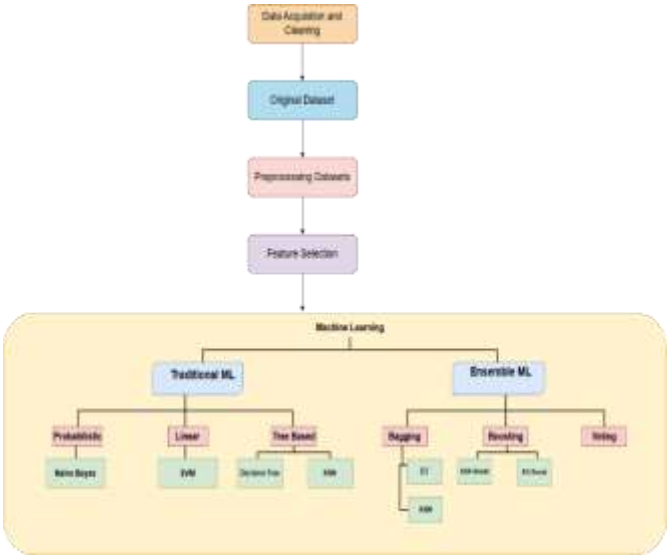

Figure1: The Architecture of Proposed Work

Common steps in preprocessing:

Data integration: Combining data from multiple sources into a unified dataset. Data normalization: Scaling features to a consistent range (e.g., between 0 and 1) to ensure fair comparison.Encoding categorical data: Converting non-numeric data into numeric formats (e.g., one-hot encoding, label encoding).

3.4 Feature Selection

Feature selection is the process of choosing the most relevant features (variables) from the dataset to improve model performance and reduce complexity. Importance: Simplifies models by focusing on significant predictors. Reduces overfitting by eliminating irrelevant or redundant features. Improves computational efficiency. Techniques: Filter methods: Select features based on statistical measures like correlation or variance (e.g., Pearson correlation, chi-squared test). Wrapper methods: Use machine learning models to test combinations of features (e.g., forward selection, backward elimination). Embedded methods: Feature selection occurs as part of the model training process (e.g., Lasso regression, decision trees)

3.5 Traditional Machine Learning

Traditional ML models are standalone algorithms that learn from data and make predictions. These models can be further classified into Probabilistic Models, Linear Models, and Tree-Based Models.

a) Probabilistic Models

Naïve Bayes: Based on Bayes' theorem with an assumption that features are independent. Works well for text classification (spam detection, sentiment analysis).

Types: *Gaussian Naïve Bayes:* Used for continuous data.

Multinomial Naïve Bayes: Used for text classification *Bernoulli Naïve Bayes:* Used for binary feature data.

b) Linear Models: Support Vector Machine (SVM): A supervised learning model used for classification and regression. Finds the best hyperplane that separates different classes Works well for high dimensional data (e.g., text classification, image recognition). Can use different kernels (Linear, Polynomial, Radial Basis Function (RBF)).

c) Tree-Based Models

Decision Tree (DT): A tree-structured model that splits data based on feature conditions.Used for both classification and regression (CART - Classification and Regression Trees).Prone to overfitting, but ensemble methods can improve it.

K-Nearest Neighbors(KNN): A non-parametric, instance-based algorithm. Classifies a new data point based on majority voting from its nearest neighbors. Works well when data has clear clusters but struggles with large datasets.

3.6 Ensemble Machine Learning

Ensemble ML combines multiple models to improve performance and robustness. There are three main types: Bagging, Boosting, and Voting.

a) Bagging (Bootstrap Aggregating)

Uses multiple models trained on different subsets of the dataset.Reduces variance and improves stability in predictions.Common bagging methods: Random Forest: An ensemble of decision trees. Bagged KNN: Applies bagging to KNN.

b) Boosting Trains weak models sequentially, giving more weight to misclassified instances.Reduces bias and improves accuracy.Popular boosting algorithms:

AdaBoost (Adaptive Boosting): Improves weak learners by focusing on mistakes. Gradient Boosting (GBM): Optimizes errors using gradient descent. XGBoost (Extreme Gradient Boosting): Faster and more efficient than GBM. LightGBM & CatBoost: More optimized versions of boosting algorithms.

c) Voting

Combines predictions from multiple models. Two types of voting: Hard Voting: Majority vote determines the final class. Soft Voting: Uses probability scores from models to decide the final prediction.

4.Result and Discussion

The implementation of machine learning techniques for health monitoring and disease prediction yielded promising results. Various models, including Decision Trees, KNN, Support Vector Machines (SVM),Naïve Bayes, XGBoost and Ada Boost were evaluated based on key performance metrics such as accuracy, precision, recall, and F1-score.

Table 1: Decision Tree Classification Report

Disease Type	**Precision**	**Recall**	**f1-score**	**Support**
Chronic	1.00	1.00	1.00	933
Normal	1.00	1.00	1.00	1028
Severe	1.00	1.00	1.00	1039
Accuracy			1.00	3000
Macro avg	1.00	1.00	1.00	3000
Weighted avg	1.00	1.00	1.00	3000

4.1 Decision Tree Classifier:

Decision Trees are a widely used supervised learning algorithm for classification and regression tasks. They work by recursively splitting the dataset into subsets based on feature conditions, creating a tree-like structure with nodes representing decisions and leaf nodes representing final classifications or predictions. Their simplicity and interpretability make them useful in medical applications, where transparency is crucial. Decision trees are easy to understand and interpret, making them useful for a wide range of applications. However, they can be prone to overfitting if not properly pruned or regularized.

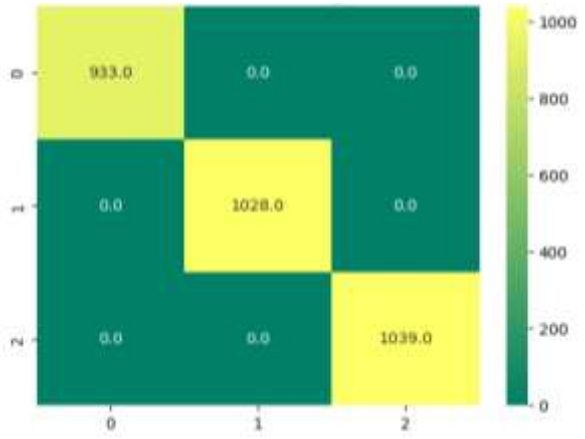

Figure 2: Decision Tree Classifier Confusion Matrix

4.2 K-Nearest Neighbors Classifier

The **k-Nearest Neighbors (KNN)** algorithm is a simple yet powerful supervised machine learning technique used for classification and regression tasks. It works by finding the **k** closest data points (neighbors) to a given input based on a chosen distance metric, such as **Euclidean, Manhattan, or Minkowski distance**. The classification is determined by a majority vote among the nearest neighbors, while regression takes the average (or weighted average) of their values. Despite its simplicity, KNN remains widely used in **pattern recognition, recommendation systems, and medical diagnosis** due to its effectiveness in many real-world applications.

Table 2: K-NN Classification Report

Disease Type	**Precision**	**Recall**	**f1-score**	**Support**
Chronic	1.00	1.00	1.00	933
Normal	1.00	1.00	1.00	1028
Severe	1.00	1.00	1.00	1039
Accuracy			1.00	3000
Macro avg	1.00	1.00	1.00	3000
Weighted avg	1.00	1.00	1.00	3000

4.3 Naïve Bayes Classifier

Naïve Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem, which assumes that features are conditionally independent given the class label. It is widely used for classification tasks such as spam detection, sentiment analysis, and medical diagnosis due to its simplicity and efficiency. Despite its strong independence assumption, Naïve Bayes often performs well in practice, especially with high-dimensional data. The algorithm calculates the probability of each class given the input features and selects the class with the highest probability.

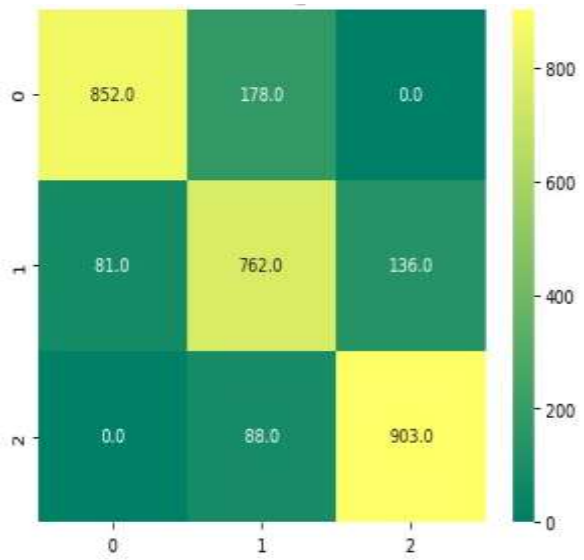

Figure 3: K-NN Classifier Confusion Matrix

4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space.

SVM aims to maximize the margin between the closest data points (support vectors) and the decision boundary, improving generalization to unseen data. Traditional machine learning techniques refer to well-established algorithms used for tasks such as classification, regression, and clustering. These include methods like Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Linear or Logistic Regression. These models rely heavily on structured data and require manual feature extraction, where relevant features must be selected and engineered before training.

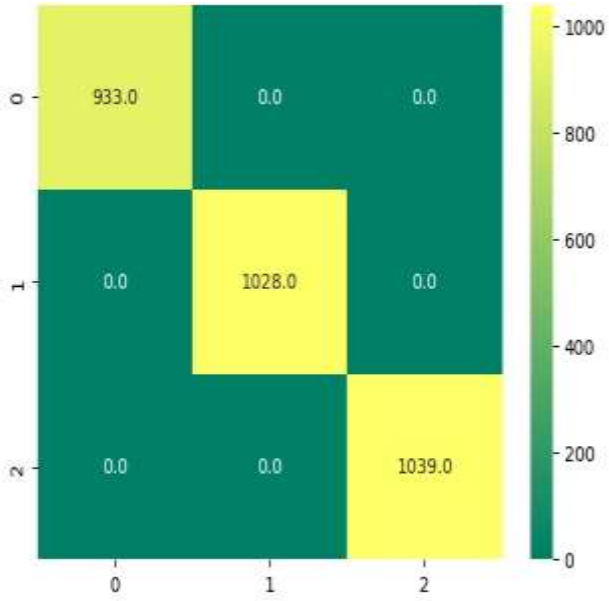

Figure 4: Naïve Bayes Confusion Matrix

4.5 XG Boost

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm based on gradient boosting that excels in both speed and performance. It builds decision trees sequentially, with each tree correcting the errors of the previous one. XGBoost includes features like regularization (L1 and L2), missing value handling, and parallelization, which enhance its efficiency and prevent overfitting[7].

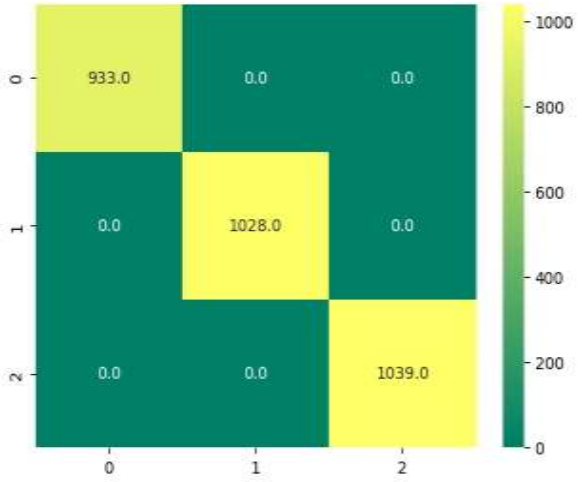

Figure 5: Support Vector Machine Confusion Matrix

4.6 XGBoost Model Training Process

Training an XGBoost model involves several key steps to optimize performance. First, the dataset is prepared by splitting it into training and testing sets. The training data is then used to build the model. An XGBClassifier or XGBRegressor is initialized with specific hyperparameters, such as the number of estimators, learning rate, and maximum tree depth, which influence model complexity and training behaviour[10].

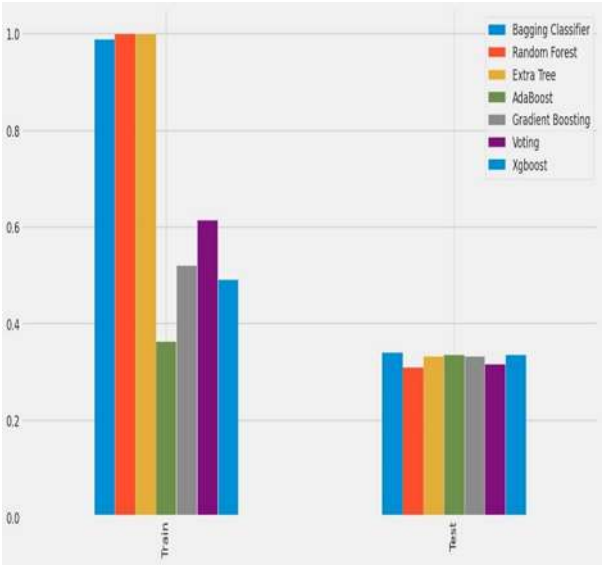

Figure 6: XGBoost Result

Table 3: Naïve Bayes Classification Report

Disease Type	Precision	Recall	f1-score	Support
Chronic	1.00	1.00	1.00	933
Normal	1.00	1.00	1.00	1028
Severe	1.00	1.00	1.00	1039
Accuracy			1.00	3000
Macro avg	1.00	1.00	1.00	3000
Weighted avg	1.00	1.00	1.00	3000

Table 4: SVM Classification Report

Disease Type	Precision	Recall	f1-score	Support
Chronic	0.83	0.91	0.87	933
Normal	0.78	0.74	0.76	1028
Severe	0.91	0.87	0.89	1039
Accuracy			0.84	3000
Macro avg	0.84	0.84	0.84	3000
Weighted avg	0.84	0.84	0.84	3000

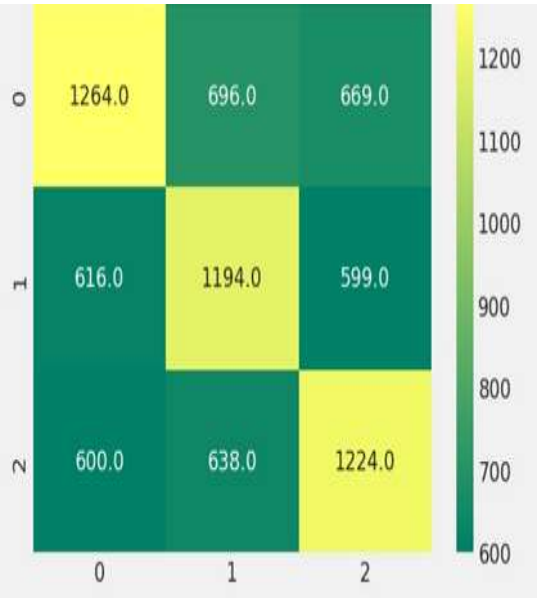

Figure 7: XGBoost Training Confusion Matrix

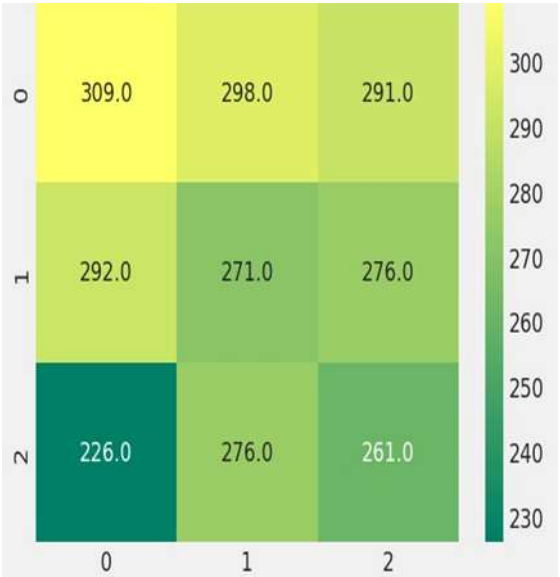

Figure 8: XGBoost Testing Confusion Matrix

During the training process, XGBoost constructs an ensemble of decision trees, where each tree corrects the errors of the previous one, improving the overall model accuracy. The model is trained iteratively, where each new tree focuses on minimizing the residual errors of the previous models, ultimately resulting in a robust model capable of generalizing well on unseen data.

4.7 XGBoost Model Testing Process

Testing an XGBoost model involves evaluating its performance on a separate test set that was not used during training. After the model has been trained, predictions are made on the test data using the .predict() method. These predictions are then compared to the true values in the test set to assess the model's accuracy or other relevant metrics, such as precision, recall, or F1-score for classification tasks, or mean squared error (MSE) for regression tasks[9].

Additionally, the model's performance may be further improved by fine-tuning hyperparameters or applying techniques such as cross-validation to ensure robust testing results.

4.8 XGBoost Evaluation Process

GridSearchCV is used to optimize hyperparameters in XGBoost by performing an exhaustive search over a parameter grid, evaluating model performance using cross-validation. It helps identify the best combination of parameters like learning rate, tree depth, and estimators to minimize the mean error (merror), which measures incorrect predictions. After training, the model's performance is tested on a separate test set to assess its generalization[8]. By minimizing merror and ensuring high accuracy on the test data, GridSearchCV helps improve both training and testing results, reducing overfitting and enhancing the model's reliability.

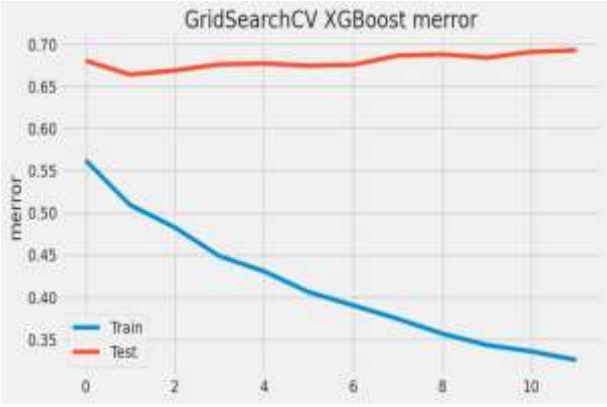

Figure 9: XGBoost Misclassification Error Analysis Plot

5.Conclusion

Machine Learning (ML) techniques play a crucial role in health monitoring and disease prediction by enabling accurate, data-driven diagnosis. Traditional models often struggle with generalization, while advanced models like XGBoost offer superior performance by effectively handling complex patterns and imbalanced data. The classification task involves 10,000 samples from five different datasets. The results showed that decision tree classification achieved an accuracy of 100% ,K-NN achieved 100%, SVM achieved 84% and naïve bayes achieved 100% .Although XGBoost reduces misclassification errors, further optimization through feature engineering and hyperparameter tuning can enhance its accuracy. Overall, ML-driven approaches, especially non-traditional models, significantly improve disease prediction, paving the way for more efficient and proactive healthcare solutions.

6.Future Enhancement

Future enhancements in health monitoring and disease prediction using ML can focus on improving model accuracy through advanced techniques like deep learning, ensemble learning, and explainable AI. Integrating real-time data from wearable devices, optimizing feature selection, and handling class imbalances can further enhance predictions. Additionally, incorporating federated learning for privacy-preserving healthcare analytics and deploying ML models in cloud-based systems can improve scalability and accessibility, leading to more precise and personalized healthcare solutions.

References

1.Forrest, Iain S., et al. "Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts." The Lancet 401.10372 (2023): 215-225.

2.Srujan, S., et al. "Skin Disease Detection using Convolutional Neural Network." International Research Journal of Engineering and Technology (IRJET) (2022).

3.Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." Scientific data 5.1 (2018): 1-9.

4.Yaseen, Gui-Young Son, and Soonil Kwon. "Classification of heart sound signal using multiple features." Applied Sciences 8.12 (2018): 2344.

5.Pal, Madhumita, and Smita Parija. "Prediction of heart diseases using random forest." Journal of Physics: Conference Series. Vol. 1817. No. 1. IOP Publishing, 2021.

6. Karthik, R., Tejas Vaichole, and Sanika Kulkarni. "Channel Attention based Convolutional Network for skin disease classification." ScienceDirect (2021).

7.Arora, Ridhi, et al. "Automated skin lesion segmentation using attention-based deep convolutional neural network." Biomedical Signal Processing and Control 65 (2021): 102358.

8. Czajkowska, Joanna, et al. "Deep learning approach to skin layers segmentation in inflammatory dermatoses." Ultrasonics 114 (2021): 106412.

9. Raihan, M., et al. "Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design." 2016 19th International Conference on Computer and Information Technology (ICCIT). IEEE, 2016.

10.K. Ashok, R. Boddu, S.A. Syed, V.R. Sonawane, R.G. Dabhade, and P.C.S. Reddy, "GAN Base feedback analysis system for indus trial IOT networks", Automatika (Zagreb), vol. 64, no. 2, pp. 259 267, 2023. http://dx.doi.org/10.1080/00051144.2022.2140391

11. M. Palumbo, B. Pace, M. Cefola, F.F. Montesano, G. Colelli, and G. Attolico, "Non-destructive and contactless estimation of chloro phyll and ammonia contents in packaged fresh-cut rocket leaves by a Computer Vision System", Postharvest Biol. Technol., vol. 189, p. 111910, 2022. http://dx.doi.org/10.1016/j.postharvbio.2022.111910

12.L. Liu, M. Shafiq, V.R. Sonawane, M.Y.B. Murthy, P.C.S. Reddy, and K.M.N.C. Reddy, "Spectrum trading and sharing in unmanned aerial vehicles based on distributed blockchain consortium system", Comput. Electr. Eng., vol. 103, p. 108255, 2022. http://dx.doi.org/10.1016/j.compeleceng.2022.108255

13. R. Nanmaran, S. Srimathi, G. Yamuna, S. Thanigaivel, A.S. Vick ram, A.K. Priya, A. Karthick, J. Karpagam, V. Mohanavel, and M. Muhibbullah, "Investigating the role of image fusion in brain tumor classification models based on machine learning algorithm for per sonalized medicine", Comput. Math. Methods Med., vol. 2022, pp. 1-13, 2022. http://dx.doi.org/10.1155/2022/7137524 PMID: 35178119

14. R. Dhanalakshmi, N.P.G. Bhavani, S.S. Raju, P.C. Shaker Reddy, D. Mavaluru, D.P. Singh, and A. Batu, "Onboard pointing error de tection and estimation of observation satellite data using extended kalman filter", Comput. Intell. Neurosci., vol. 2022, pp. 1-8, 2022. http://dx.doi.org/10.1155/2022/4340897 PMID: 36248921

15.L.E. Doyle, J.R. Loeb, N. Ekramirad, D. Santra, and A.A. Adedeji, "Non-destructive classification and quality evaluation of proso mil let cultivars using NIR hyperspectral imaging with machine learn ing", 2022 ASABE Annual International Meeting, 2022, p. pp. 1 http://dx.doi.org/10.13031/aim.202200944

16. P. Reddy, and A. Sureshbabu, "An adaptive model for forecasting seasonal rainfall using predictive analytics", Int J Intell Eng Syst, vol. 12, no. 5, pp. 22-32, 2019. http://dx.doi.org/10.22266/ijies2019.1031.03

17. R. Sabitha, A.P. Shukla, A. Mehbodniya, and L. Shakkeera, "A fuzzy trust evaluation of cloud collaboration outlier detection in wireless sensor networks", Ad Hoc Sens. Wirel. Netw., vol. 53, no. 3/4, pp. 165-188, 2022.

18 J.F.I. Nturambirwe, and U.L. Opara, "Machine learning applica tions to non-destructive defect detection in horticultural products", Biosyst. Eng., vol. 189, pp. 60-83, 2020. http://dx.doi.org/10.1016/j.biosystemseng.2019.11.011

19. Y. Hou, X. Cai, P. Miao, S. Li, C. Shu, P. Li, W. Li, and Z. Li, "A feasibility research on the application of machine vision technology in appearance quality inspection of Xuesaitong dropping pills", Spectrochim. Acta A Mol. Biomol. Spectrosc., vol. 258, p. 119787, 2021

20.Mrutyunjaya M.S and Harish Kumar K.S.Non-destructive Machine Vision System Based Rice Classification Using Ensemble Machine Learning Algorithms. Volume 17, Issue 5, 2024