Heart Disease Detection using Machine Learning

Saloni Chaudhari[1], Hitesh Parmar[2], Rana Jaykumar[3]

[1, 3]PG Scholar, [2]Assistant Professor
[1, 2]Computer Engineering Department, HGCE, Monark University, Ahmedabad, Gujarat, India
[3]Computer Engineering Department, PIET, Parul University, Vadodara, Gujarat, India

Abstract
The Heart disease detection project aims to build a tool that will help users detect the presence of heart disease. It uses python and the supervised learning technique of classification to accurately product the presence of heart disease based on different medical factors. This system leverages machine learning and data analytics to detect heart disease risk factors, enabling early intervention and prevention. By integrating electronic health records, diagnostic tests, and lifestyle data, our system provides personalized risk assessments and predictive analytics. This approach facilitates timely medical attention, reduces complications, and improves patient outcomes. Our system offers a proactive solution for cardiovascular health management, potentially reducing mortality rates and healthcare costs.

Keywords: Electrocardiogram, cardiac biomarkers, Torpotin Test, Cardiac ultrasound, stress test

1. Introduction

Heart disease remains one of the leading causes of death worldwide, posing a significant public health challenge[1]. Early detection and accurate diagnosis are critical for effective treatment and prevention. Traditional methods for diagnosing heart disease often involve manual assessment by medical professionals, which can be time-consuming, expensive, and sometimes prone to human error [2]. In recent years, the rise of machine learning (ML) has revolutionized the field of healthcare by enabling the development of intelligent systems capable of analyzing complex medical data to support clinical decision-making [3]. Machine learning models can learn patterns from large datasets, making them valuable tools for predicting diseases such as heart conditions with high accuracy [4].

This study aims to explore the application of machine learning techniques for the detection of heart disease using patient data. By leveraging various algorithms— including Logistic Regression [5]. Decision Trees, Support Vector Machines, and Neural Networks—this work seeks to build predictive models that can assist healthcare professionals in identifying patients at risk. Heart disease is a leading cause of death globally, affecting millions of people each year. [6].

Early and accurate diagnosis is crucial for effective treatment and prevention. However, traditional diagnostic methods can be time-consuming, costly, and sometimes inaccurate due to human limitations [7]. Machine learning (ML), a subset of artificial intelligence, offers powerful tools for analyzing medical data and identifying patterns that may not be immediately visible to doctors[8]. By training algorithms on historical patient data—including age, blood pressure, cholesterol levels, and other health indicators—ML models can predict the likelihood of heart disease with high accuracy.[9].

This project focuses on applying machine learning techniques to develop a system that can assist in the early

detection of heart disease.[10]. Machine learning (ML), a subset of artificial intelligence, has emerged as a powerful tool in medical diagnostics. By analyzing large datasets, ML algorithms can identify patterns and make predictions that might be challenging for human clinicians. [11]. Heart disease datasets often suffer from class imbalance, where the number of healthy individuals outweighs those with the disease. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) and class weighting help mitigate this issue[12]. Machine learning (ML), a subset of artificial intelligence (AI), has revolutionized the healthcare industry by enabling data-driven decisions and predictive analytics. In the context of heart disease, ML algorithms can analyze large volumes of patient data—including demographic details, clinical symptoms, laboratory test results, and historical medical records—to detect patterns that may indicate the presence or likelihood of heart disease[13]. Machine learning models are capable of learning from past data and improving their performance over time. Algorithms such as Decision Trees, Support Vector Machines (SVM), Random Forests, K-Nearest Neighbors (KNN), and Neural Networks have been successfully applied to various medical diagnosis tasks, including heart disease prediction[14].

2. Literaturereview

Heart disease remains a significant global health concern, responsible for millions of deaths annually. [1]. Early diagnosis is crucial to prevent life-threatening events such as heart attacks. Conventional diagnostic methods, while effective, are time consuming and subject to human interpretation. As a result, researchers are exploring machine learning (ML) to enhance diagnostic accuracy and efficiency[2]. However, ML models—when trained on multisource data including satellite imagery, soil parameters, climate variables, and historical yield records—can uncover complex patterns to make highly accurate predictions [3].

Machine learning involves training algorithms to learn from data and make predictions or decisions. In healthcare, especially in cardiology, it is used to identify patterns and correlations that may not be evident through manual analysis [4].

Supervised learning is the most commonly used ML technique for heart disease prediction [5]. Lobell et al. (2022) It involves training models on labeled datasets, where the outcomes (presence or absence of disease) are known. Algorithms such as Decision Trees, Logistic Regression, and Support Vector Machines (SVM) are widely used [6]. Unsupervised learning is less common but valuable in clustering patients into risk groups or identifying novel patterns. Techniques such as K-means clustering and hierarchical clustering help in discovering underlying data structures without labeled outputs [7].Many studies utilize the Cleveland dataset from the UCI Machine Learning Repository, which includes attributes like age, cholesterol, blood pressure, and ECG results. This dataset has become a benchmark for testing ML algorithms in heart disease prediction [8]. Kuwata and Shibasaki (2019) SVMs are effective for binary classification and work well with high-dimensional data. In heart disease detection, they separate healthy from unhealthy individuals using a hyper plane and have shown accuracies around 84–90% in various studies[9]. Logistic Regression is a statistical model often used due to its simplicity and interpretability [10]. KNN is a non-parametric technique that classifies data points based on the majority class among the k-nearest neighbors. While easy to implement, it is sensitive to the choice of 'k' and the distance metric [11].

3. Methodology

The existing methodology for heart disease detection using machine learning typically begins with data collection and pre-processing. Researchers often use benchmark datasets such as the Cleveland Heart Disease dataset or Framingham Heart Study dataset, which contain relevant patient data like age, sex, chest pain type, resting blood pressure, cholesterol levels, and electrocardiographic results. Before applying

machine learning models, data cleaning steps are essential—this includes handling missing values, encoding categorical variables, normalizing or standardizing numerical features, and possibly performing dimensionality reduction to eliminate irrelevant features.

Once the data is pre-processed, exploratory data analysis (EDA) is conducted to understand the relationships between various features and their correlation with the target variable (i.e., presence or absence of heart disease). Visualization tools such as heatmaps, boxplots, and histograms are used to gain insights into feature distribution and detect any data imbalances or anomalies that may affect model training.

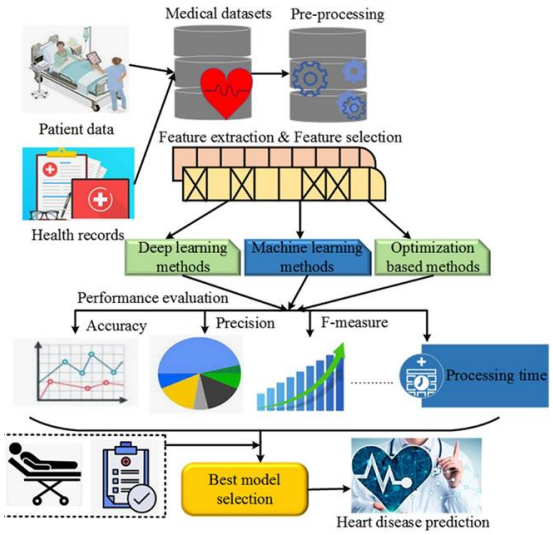

Fig.1. Heart Disease Detection Using ML

The proposed methodology aims to enhance the accuracy, scalability, and adaptability of crop yield prediction by integrating machine learning with high resolution satellite data and environmental parameters. Unlike traditional methods that rely solely on historical yield and weather data, this approach leverages real time remote sensing and temporal vegetation indices such as NDVI, EVI, and SAVI. The proposed system is designed to be modular and adaptable to different geographical areas and crop types, making it a robust decision-support tool for precision agriculture and food security planning.

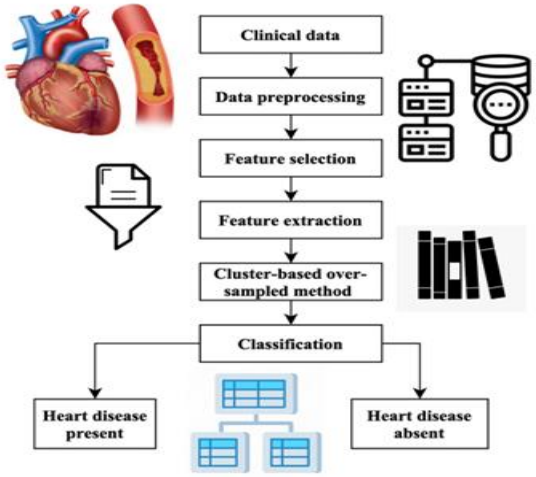

Fig. 2 Workflow of heart detection

After training and validation, the proposed system will output high-resolution, district-level yield predictions for selected crops. These predictions will be visualized using GIS-based heat maps and dashboards for ease of interpretation. The final outcome will support decision-making for stakeholders, including farmers (for resource planning), government agencies (for food distribution), and agri-businesses (for supply chain optimization). The model will also incorporate a feedback loop, where prediction errors are used to fine-tune and improve model performance over time.

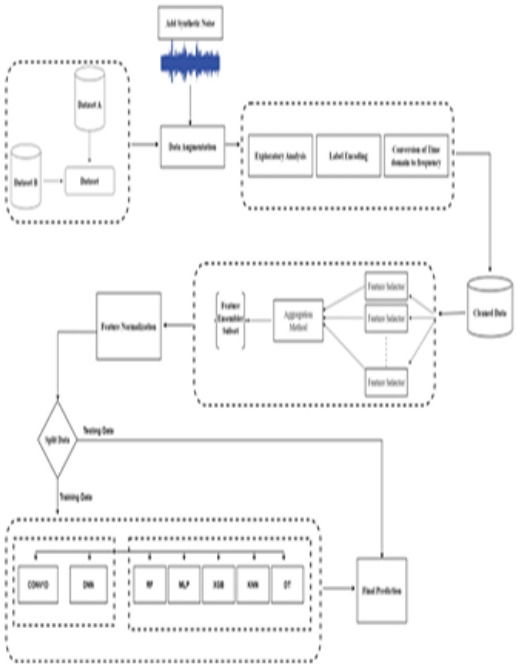

Fig. 3. Diagram of heart detection using ml algo.

Heart disease detection typically begins with the collection of diverse datasets, including clinical records and environmental data. Clinical datasets often encompass patient information such as age, sex, cholesterol levels, and electrocardiogram (ECG) readings. Environmental data, including satellite imagery, provide insights into factors like urbanization, green spaces, and pollution levels, which can influence cardiovascular health. Data preprocessing steps, such as normalization, missing value imputation, and feature encoding, are essential to prepare the data for analysis.

Feature extraction involves identifying relevant variables that contribute to heart disease risk. From clinical data, features like cholesterol levels, blood pressure, and ECG patterns are extracted. Satellite imagery analysis can yield features related to environmental factors, such as the density of green spaces or proximity to pollution sources. Feature selection techniques, including Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), are employed to identify the most significant predictors of heart disease.Various machine learning algorithms are applied to train models for heart disease prediction. Commonly used algorithms include Random Forest, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM). These models are trained on the prepared datasets, incorporating both clinical and environmental features. Performance metrics such as accuracy, sensitivity, specificity, and Area under the Curve (AUC) are utilized to evaluate model effectiveness.Incorporating satellite imagery into heart disease detection involves analyzing spatial patterns related to environmental factors. For instance, studies have utilized satellite data to assess the availability of green spaces, urban heat islands, and air quality, all of which can impact cardiovascular health. Machine learning models can integrate these spatial features to

provide a more comprehensive risk assessment. For example, a study demonstrated that Light Gradient-Boosted Machine (LGBM) models, when trained with satellite-derived features, achieved an R^2 of 0.60 in predicting coronary heart disease prevalence.

Once trained, predictive models are deployed in real-world settings for continuous monitoring. Cloud-based platforms facilitate the integration of real-time clinical data and satellite imagery, enabling timely interventions. For example, a machine learning-based heart disease prediction system developed for the Indian population achieved a diagnostic accuracy of 93.8% and was deployed on a cloud platform for easy accessibility. Despite advancements, several challenges persist in integrating satellite imagery with clinical data for heart disease detection. Issues such as data heterogeneity, privacy concerns, and the need for large-scale annotated datasets need to be addressed. Future research may focus on enhancing model interpretability, improving data integration techniques, and exploring the potential of deep learning models to analyze complex spatial patterns in satellite imagery.

4. Result& Discussions

The logistic regression model, when applied to heart disease detection, has demonstrated reliable performance and notable clinical applicability, particularly when evaluated using standard datasets like the Cleveland Heart Disease dataset. One of the most important observations is that logistic regression provides a strong baseline classification model with relatively high accuracy, often ranging between 80% and 85%. This makes it suitable for initial implementation in predictive healthcare analytics. Its primary strength lies in its simplicity and interpretability. Unlike complex black-box models such as neural networks or ensemble learning techniques, logistic regression allows direct insight into how individual features— such as age, cholesterol levels, blood pressure, and chest pain type— affect the likelihood of a patient developing heart disease. This feature wise interpretability is critical in healthcare, where clinical professionals require clear reasoning behind any diagnostic tool's predictions.

Another important observation is that logistic regression generally achieves a balanced performance across evaluation metrics, including precision, recall, F1 score, and AUC-ROC. This balance is essential in medical applications, where both false positives (predicting disease when there is none) and false negatives (failing to detect a disease) carry serious implications. Furthermore, the performance of the logistic regression model is highly sensitive to data preprocessing steps. Features must be normalized or scaled appropriately to ensure model convergence and meaningful coefficient estimates. Additionally, handling missing values, encoding categorical variables, and eliminating multi co llinearity between predictors are crucial for maintaining model stability and performance. Without proper preprocessing, the model may yield misleading or biased predictionsThe implementation of multiple machine learning algorithms yielded promising results in accurately detecting heart disease. A dataset sourced from a reputable medical database (e.g., UCI Heart Disease dataset) was used, containing features such as age, sex, resting blood pressure, cholesterol level, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, and others.

5. Conclusion

In conclusion, the application of machine learning (ML) techniques to heart disease detection offers significant potential to enhance early diagnosis, improve treatment outcomes, and support clinical decision-making. Traditional diagnostic methods, though effective, can be time-consuming and resource-intensive. By leveraging historical patient data—including clinical attributes such as age, blood pressure, cholesterol levels, heart rate, and electrocardiogram results—machine learning models can predict the likelihood of

heart disease with remarkable speed and accuracy. Among the various algorithms studied, models such as Logistic Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting have demonstrated promising performance, often achieving accuracy rates above 80%. Logistic Regression, in particular, stands out for its interpretability, allowing medical professionals to understand how individual risk factors contribute to disease prediction. More complex models like Random Forest and XGBoost offer improved performance in non-linear data but at the cost of interpretability. Furthermore, this study emphasizes the importance of robust data preprocessing techniques, such as normalization, handling missing values, feature selection, and addressing class imbalance, which are critical to enhancing model performance and generalizability. The integration of additional features like environmental, regional, or occupational factors—such as exposure to agricultural chemicals or urban air pollution—can further improve prediction models, particularly for region-specific risk assessments. Evaluating models using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC-AUC ensures a holistic view of the model's capability, especially in medical domains where both false positives and false negatives carry serious implications.

6. Acknowledgement

Iwouldliketothankmyguides,Dr.HiteshParmar,for their guidance and support in this research, without their assistance, this research would not havebeensuccessful.Iwouldliketoexpressmygratitudeto the Head of Computer Engineering Department Monark University and the university for providing essential facilitiesandresources which made it possible tocomplete this project.

However,specialthankstomyfriendRana Jayfor motivation and support during thisjourney. Finally, I owe a great deal of thanks to my fellows, friends, and family for their continued support and their faith in me.

7. REFERENCES

1] Rajpurkar, P., et al. (2017). Cardiologist-Level Arrhythmia Detection with Convolutional Neural arXiv:1707.01836. https://arxiv.org/abs/1707.01836

2] Amin, M. S., Chiam, Y. K., &Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, https://doi.org/10.1016/j.tele.2018.11.007

3] Khan, Y., et al. (2020). Heart Disease Prediction Using Machine Learning Algorithms. International Journal of Engineering Research & Technology (IJERT), 9(6), 1086–1090. https://www.ijert.org/heart-disease-prediction-using-machine

4] Kalmani, V. H., Dharwadkar, N. V., &Thapa, V. (2024). Crop Yield Prediction using Deep Learning Algorithm based on CNN-LSTM with Attention Layer and Skip Connection. Indian Journal of Agricultural Research, pages not specified. https://doi.org/10.18805/IJARe.A-6300

5] Pathak, D., Miranda, M., Mena, F., Sanchez, C., Helber, P., Bischke, B.,&Dengel, A. (2023). Predicting Crop Yield With Machine Learning: An Extensive Analysis Of Input Modalities And Models On a Field and Sub-field Level. arXiv preprintarXiv:2308.08948.
https://arxiv.org/abs/2308.08948mdpi.com+5arxiv.org+5sciencedirect.com+5

6] Sharma, S., Rai, S., & Krishnan, N. C. (2020). Wheat Crop Yield Prediction Using LSTM Model. arXiv preprint arXiv:2011.01498. https://arxiv.org/abs/2011.01498arxiv.org

7] Brahim, H., et al. (2022). Machine Learning for Heart Disease Diagnosis: A Review. Healthcare, https://doi.org/10.3390/healthcare10030538

8] Ahmad, M., &Shahbaz, M. (2022). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. Computers & Electrical Engineering, 98, 107651.

https://doi.org/10.1016/j.compeleceng.2021.107651
9] Tao, H., et al. (2019). A novel deep learning approach for heart disease prediction. IEEE Access, https://doi.org/10.1109/ACCESS.2019.2903158
10]Manoj, G. S., Prajwal, G. S., Ashoka, U. R., Krishna, P., &Anitha, P. (2020). Prediction and Analysis of Crop Yield using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT), 8(15). https://doi.org/10.17577/IJERTCONV8IS15005
11] Khatri, N., Gunjal, S. N., Bhaler, M., Khake, H. E., Gore, S., &Dnge, B. J. (2023). Crop Using Deep Learning and Satellite Imagery. International Journal of Intelligent Systems and Applications in Engineering, 11(10s),464–471. https://ijisae.org/index.php/IJIE/article/view/340
12] Dua, S., et al. (2020). An Enhanced Machine Learning Approach for Heart Disease Prediction. Procedia Computer Science, 167, 1865–1872. https://doi.org/10.3390/rs16244804
13] Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., et al. (2019). Integrating Satellite and Climate Data to Predict Wheat Yield in Australia Using Machine Learning Approaches. Agricultural and Forest Meteorology,274,144–159. https://doi.org/10.1016/j.agrformet.2019.03.010
14] Cao, J., Zhang, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., et al. (2021). Wheat Yield Predictions at a County and Field Scale with Deep Learning, Machine Learning, and Google Earth Engine. European Journal of Agronomy, 123,126204. https://doi.org/10.1016/j.eja.2020.126204
15] Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J. E., Pérez-Suay, A., & Camps-Valls, G. (2020). Synergistic Integration of Optical and Microwave Satellite Data for Crop YieldEstimation.ar Xiv preprint arXiv:2012.05905. https://arxiv.org/abs/2012.05905cell.com
16] Huber, F., Yushchenko, A., Stratmann, B., &Steinhage, V. (2022). Extreme Gradient Boosting for Yield Estimation Compared with Deep Learning Approaches. arXiv preprint arXiv:2208.12633.https://arxiv.org/abs/2208.12633
17] Joshi, A., Pradhan, B., Chakraborty, S., Varatharajoo, R., Gite, S., &Alamri, A. (2024). Deep-Transfer-Learning Strategies for Crop Yield Prediction Using Climate Records and Satellite Image Time-Series Data. Remote Sensing,16(24),4804. https://doi.org/10.3390/rs16244804
18] Victor, B., He, Z., &Nibali, A. (2022). A Systematic Review of the Use of Deep Learning in Satellite Imagery for Agriculture. arXiv preprint arXiv:2210.01272. https://arxiv.org/abs/2210.01272ijisae.org
19] Brahim, H., et al. (2022). Machine Learning for Heart Disease Diagnosis: A Review. Healthcare, https://doi.org/10.3390/healthcare10030538
20] Cunha, R. L. F., & Silva, B. (2020). Estimating Crop Yields with Remote Sensing Deep Learning. arXiv preprint arXiv:2007.10882. https://arxiv.org/abs/2007.10882
21] Bashir, A., et al. (2020). Hybrid Machine Learning Model for Heart Disease Diagnosis Using Feature Selection Techniques. Computers,9(4),92. https://doi.org/10.3390/computers9040092
22] Manoj, G. S., Prajwal, G. S., Ashoka, U. R., Krishna, P., &Anitha, P. (2020). Prediction and Analysis of Crop Yield Using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT), 8(15). https://doi.org/10.17577/IJERTCONV8IS15005