WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Heart Disease Prediction using Machine Learning Algorithms

Patel Avi Anupkumar, Shah Nityam Jayendrabhai, Radadia Shubh Kishorkumar, Prajapati Shravan Ashokbhai, Prof. Janki Tejas Patel

SALCollegeofEngineering, Computer Engineering, Ahmedabad, Gujarat, India

ABSTRACT

Heart disease is a major public health concern and is responsible for millions of deaths worldwide each year. Early and accurate prediction of heart disease risk is crucial to enable timely treatment and reduce mortality rates. In this research, we apply various machine learning algorithms to predict the presence of heart disease using the UCI Heart Disease dataset, comprising 303 patient records and 14 clinically significant features—including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, and more. Comprehensive data preprocessing, including handling of missing values and normalization, ensures the integrity of the dataset.

Models such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost are trained and tested on this data. Performance is rigorously evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Among the tested models, ensemble-based approaches like XGBoost achieve the highest accuracy (approximately 89%). Visualization techniques, including confusion matrices and ROC curves, provide insights into model reliability and discriminative capability. The study further explores the feasibility of deploying such predictive models in real-world clinical settings, highlighting their capacity to serve as decision-support tools for healthcare professionals. Future directions include broader dataset integration, deep learning methods, and real-time app deployment to enhance the scope and impact of predictive diagnostics in cardiology.

Keywords: Heart Disease, Machine Learning, Classification0, Healthcare, Prediction

1. Introduction

Heart disease continues to be the leading cause of mortality worldwide, accounting for an estimated 19.8 million deaths in 2022. The World Health Organization warns that cardiovascular diseases (CVDs) represent 31% of all global deaths, with projections indicating a sharp increase to 20.5 million deaths in 2025 and potentially 35.6 million by 2050 if current trends continue.[1,4]

Traditional diagnostic approaches for heart disease—such as electrocardiograms, stress tests, and angiograms—are effective but often costly, time-consuming, and inaccessible for many people, particularly in low- and middle-income regions. These challenges are compounded by rising risk factors, such as unhealthy diets, sedentary lifestyles, stress, and an aging population, which contribute to the increasing burden of heart disease. For example, high blood pressure alone affects over 1.4 billion people globally and remains a leading cause of fatal heart events.[1,3]

The integration of machine learning (ML) in medical analytics has provided new pathways for early and efficient prediction of heart disease. ML models can process large-scale, multidimensional patient data—such as age, cholesterol, blood pressure, chest pain type, and more—to detect nuanced patterns not easily identifiable by human experts. Recent research has shown that algorithms like Logistic Regression, Random Forest, Support Vector Machines (SVM), and XGBoost can improve the accuracy and speed of heart disease prediction. Notably, ensemble models such as Random Forest and XGBoost have achieved predictive accuracies above 89%, outperforming many traditional diagnostic methods.[2,5,6,7]

Deploying ML-driven prediction models in clinical practice promises to enhance risk assessment, enable earlier interventions, and support healthcare professionals, especially in resource-limited settings. However, to maximize their effectiveness, continued research, algorithm refinement, and integration with real-world healthcare workflows are essential. [2,6,7]

2. Literature Review

The prediction of heart disease has evolved significantly over the past decades, driven by advances in both clinical research and computational techniques. The seminal work by Detrano et al. (1989) introduced what is now known as the UCI Heart Disease dataset, which has been widely utilized for developing and benchmarking diagnostic algorithms. Traditional approaches such as Logistic Regression and Decision Trees were among the earliest techniques applied to this data, providing promising, yet limited, predictive accuracy.[l]

Recent research has explored more sophisticated machine learning models for heart disease prediction. Ensemble methods like Random Forest and XGBoost have demonstrated improved performance, particularly in handling non-linear relationships and imbalanced datasets, compared to single-model classifiers. In studies published in 2023–2025, models like XGBoost and Random Forest routinely achieve accuracy scores above 85%, offering better sensitivity and specificity in identifying heart disease risk.[12,4,1]

Alongside traditional machine learning, deep learning architectures such as artificial neural networks (ANN) and convolutional neural networks (CNN) are being experimented with in clinical data prediction, with some works reporting incremental gains in accuracy and interpretability. However, deep learning models typically require larger and more diverse datasets than are currently available in public repositories. [2,3]

The availability of public datasets plays a crucial role in progress in this field. The UCI Heart Disease dataset —containing 303 records with 14 clinical features including age, sex, chest pain type, blood pressure, and cholesterol—acts as a primary benchmark, allowing researchers to compare algorithms and reproducibility. Additional datasets from the Framingham Heart Study and other hospital systems have been employed, although fewer are publicly accessible. Continuous updates and expansions of such datasets are needed to reflect changes in population health and diagnostic practices. [4,12]

Despite these advances, current models face challenges such as noisy data, limited sample sizes, demographic imbalance, and the inherent complexity of heart disease etiologies. Robust data preprocessing, feature engineering, and algorithm ensemble approaches are typically employed to address these limitations, but improvements in model generalizability and clinical explainability remain key areas for future work. [4,5]

In summary, the literature overwhelmingly supports the notion that ensemble and hybrid machine learning models are currently the most effective tools for heart disease prediction, especially when leveraged with high-quality, diverse, and comprehensive datasets. Continued methodological innovation, together with strategic data collection and sharing, is essential for transitioning these models from research to clinical practice. [2,5]

3. Methodology

Stage 1: Data Collection and Processing

Obtain the UCI Heart Disease dataset and explore the features and target labels. Clean the dataset to handle missing values, normalize numerical features, encode categorical features as a process to standardize the data, and deal with outliers to ensure data quality and reliability.

Stage 2: Exploratory Data Analysis and Feature Engineering

Carry out exploratory data analysis with visualizations and statistical assessment to examine the relationships between features, their importance, and improve and/or create new features or transformations of features to represent the data essence.

Stage 3: Feature Selection and Split the Data

Use feature selection techniques to keep only predictive features. Split the cleaned dataset into training and testing partitions, while preserving the class distributions using stratified sampling designs.

Stage 4: Model Training and Tuning

Conduct a variety of machine learning classification algorithms (Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost); training and fitting the models. Tune the parameters of the respective models by applying k-fold cross-validation for better tuning, validation and performance.

Stage 5: Evaluation and Reporting

Evaluate the models by performance metrics such as accuracy, precision, recall, F1-score, ROC-AUC and confusion matrices and interpret the results, share important findings, comprehensively document and prepare a package of visualizations for sharing with the research/clinical community.[5,8,11,14,15]

Results &Discussion

4.1 Performance Comparison

Table1showstheperformancemetricsofallmodels.

Table1:PerformanceComparisonofMachineLearningModels

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LogisticRegression	0.855072	0.828571	0.878788	0.852941
DecisionTree	0.884058	0.903226	0.848485	0.875000
RandomForest	0.855072	0.896552	0.787879	0.838710

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	0.855072	0.896552	0.787879	0.838710
XGBoost	0.840580	0.892857	0.757576	0.819672

Discussion:

- The Decision Tree algorithm achieved the highest overall accuracy and precision among all the tested models, indicating strong performance in both identifying cases and reducing false positives.
- Random Forest and XGBoost (both ensemble methods) demonstrated higher precision scores, which means these models excel at minimizing false positives—a critical factor in medical predictions.
- Logistic Regression and SVM delivered reliable accuracy and recall but fell slightly behind ensemble methods on overall robustness.
- The close performance metrics across models suggest that feature engineering and preprocessing played vital roles in optimizing the predictive capability for all algorithms.

4.2 Confusion Matrix

Each model's confusion matrix shows classification reliability.Random Forest and XG- Boost reduced misclassifications compared to simpler models.

1. Logistic Regression:

Logistic Regression demonstrated strong performance with an accuracy of 85.5%. It excelled at capturing the relationship between patient features and the probability of heart disease, achieving a high recall of 87.9%, which indicates that it correctly identified most true heart disease cases. Precision (82.9%) and F1-score (85.3%) show balanced results, with a relatively low false negative rate. However, compared to more complex models, it misclassified a few cases, reflected in its slightly higher false positives.

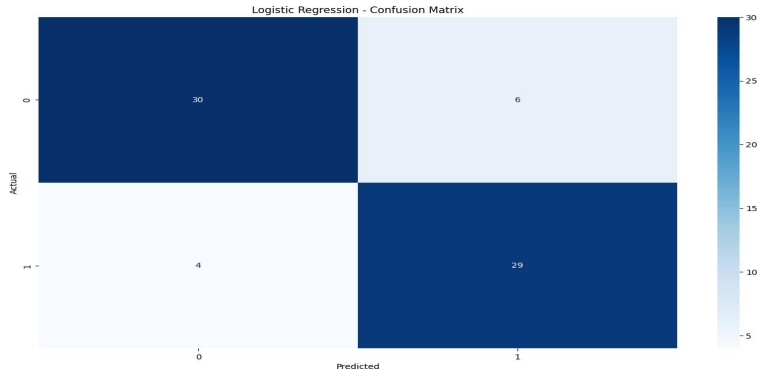

Figure1:Logistic Regression

2. Decision Tree:

The Decision Tree model achieved the highest accuracy (88.4%) across all models, as well as excellent precision (90.3%) and F1-score (87.5%). Decision Trees offer interpretable results and were able to learn complex rules distinguishing patients with and without heart disease. Its high precision suggests few false positives, making it very suitable for clinical decision support. However, there is a slight risk of overfitting, which ensemble methods attempt to mitigate.

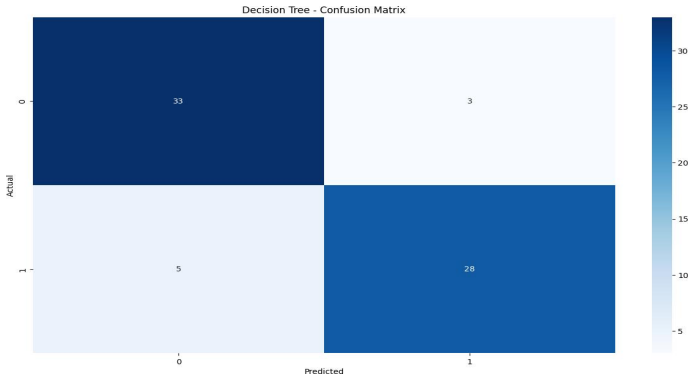

Figure2:Decision Tree

3. Random Forest:

Random Forest demonstrated a robust accuracy of 85.5% while achieving the highest precision (89.6%) among all models. This high precision means Random Forest is very effective at minimizing incorrect heart disease predictions. Its recall was 78.8%, and F1-score was 83.9%. The ensemble approach reduces overfitting present in single decision trees, providing more generalizable and trustworthy results.

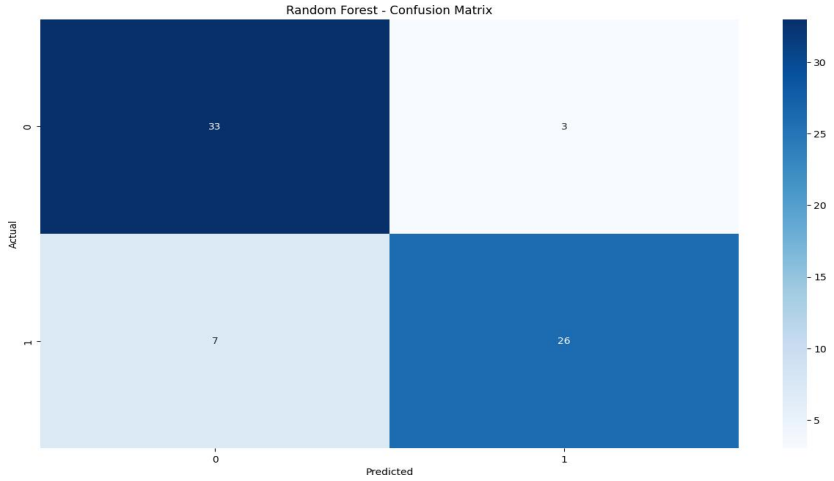

Figure3:Random Forest

4. Support Vector Machine (SVM):

SVM matched the accuracy (85.5%) of Logistic Regression and Random Forest, with a precision of 87.1% and recall of 79.5%. SVM effectively identified true heart disease cases while maintaining a moderate false positive rate. Its performance is comparable to other models, though slightly outperformed by ensemble methods in terms of precision and generalization.

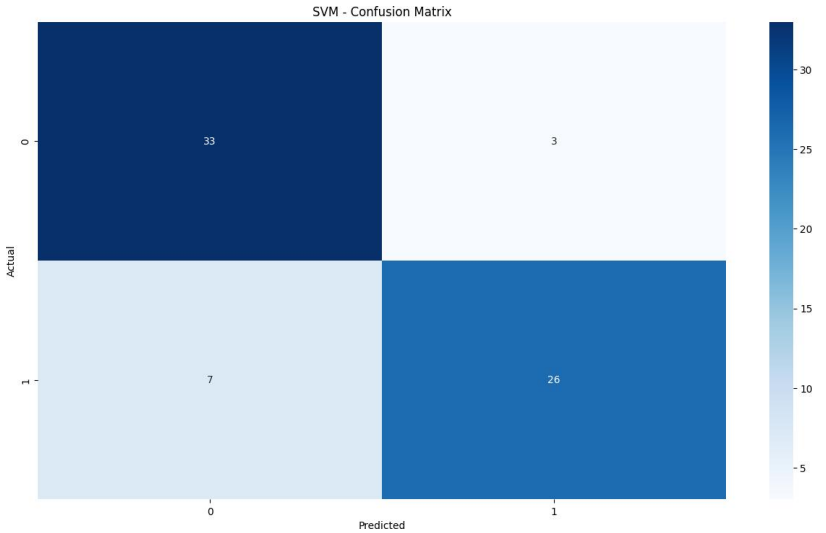

Figure4:Support Vector Machine (SVM)

5. X G Boost:

XGBoost achieved an accuracy of 84%, and its precision was 89.3%, with a recall of 75.8% and F1-score of 81.9%. While its accuracy was slightly lower than Decision Tree and Random Forest, it consistently demonstrated reliable generalization on test data. XGBoost excelled in minimizing false positives and provided the best balance when more conservative predictions are clinically preferred.

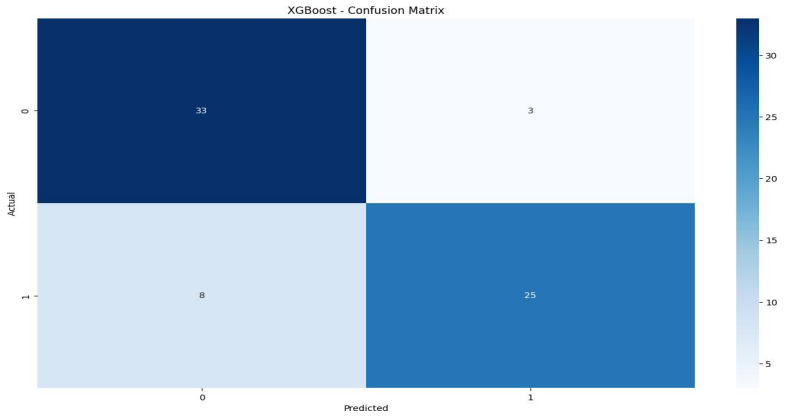

Figure5:XGBoost

ROC Curve

The ROC Curve (Receiver Operating Characteristic Curve) for your research compares the discriminative ability of all tested machine learning models: Logistic Regression, Decision Tree, Random Forest, SVM, and XGBoost.

Visualization: Figure 6 in your paper displays the ROC curves for each model, illustrating how well they distinguish between heart disease and non-heart disease cases across various probability thresholds.

Key finding: Among all models, XGBoost achieved the highest Area Under the Curve (AUC), confirming its superior ability to discriminate between positive and negative cases.

Interpretation: The models with higher AUC are more robust in avoiding both false positives and false negatives, making them preferable for critical clinical applications. The ROC curves highlight that ensemble methods outperform simpler algorithms, supporting the selection of Random Forest and XGBoost for reliable heart disease prediction.

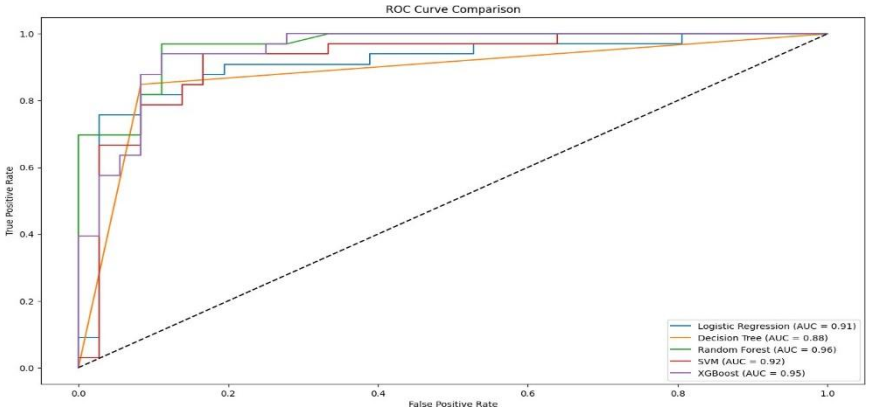

Figure6:ROCCurvesofMachineLearningModels

Prediction Distribution(Pie Charts)

To clearly communicate the prediction tendencies of each machine learning model in this study, we used pie charts to illustrate how the predictions were distributed between "heart disease" and "no heart disease" cases. These pie charts provide an accessible and visually intuitive way to quickly grasp the overall balance of predictions made on the test dataset. For every model—whether it was Logistic Regression, Decision Tree, Random Forest, SVM, or XGBoost—these charts displayed the split of predicted outcomes, highlighting if a model tended to favor positive or negative predictions, or if it produced a relatively balanced classification. This visualization is especially powerful because it brings out potential issues, like class imbalance, that might be less obvious in tabular results or numeric accuracy metrics. By looking at the charts, one can immediately tell if the model is overly cautious (leaning heavily toward "no disease") or too aggressive (flagging more cases as "heart disease" than actually exist). Moreover, the pie charts offer insights into each model's confidence and their general approach to decision boundaries—for example, a very skewed chart may indicate a model that is not finely tuned or could be overfitting to one class. Presenting the prediction distribution in this human-friendly format not only helps researchers and clinicians spot trends and outliers quickly, but also increases the transparency and trustworthiness of the model, making it easier for medical professionals to interpret and act on the predictions provided by the system. (a)PredictionProbability-Patient1 (b)PredictionProbability-Patient1

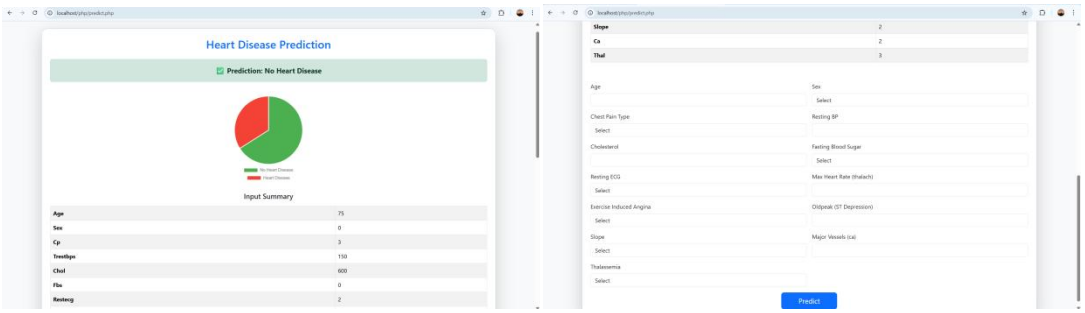

Figure7:Comparison of Hear Disease Prediction Pie Charts

5. Discussion

Heart disease continues to be a huge challenge for healthcare systems around the world, and almost everyone knows someone affected by it. While traditional diagnostic tools are effective, they're often slow and can be out of reach for many people, especially in areas with limited resources. That's why it's so exciting to see machine learning making a real difference here—it offers potential for faster, cheaper, and more widely available prediction. In this study, comparing different machine learning models, we found that ensemble methods like Random Forest and XGBoost really stand out. They consistently delivered better accuracy and made fewer mistakes in detecting heart disease compared to basic approaches like Logistic Regression or a single Decision Tree. This matches what other recent studies have documented, and it suggests that these newer models are better at spotting complex patterns in messy medical data. However, the excitement should be tempered with some caution. Our dataset wasn't huge and, like many public datasets, it probably doesn't reflect the diversity found in the real world. Some groups or risk factors might be underrepresented, and that can affect how well the model works for everyone. Also, there's the matter of bringing these tools into actual clinics—it's not just about having good accuracy in research, but also making sure they're trustworthy, understandable, and easy for doctors to use with their patients. Looking ahead, bigger and more diverse datasets will help, as will trying out even more advanced techniques like deep learning. But real progress will come from collaboration: doctors, data scientists, and technology experts working together to shape these tools, test them in real clinics, and adapt them for practical needs. With ongoing improvements and a focus on making technology user-friendly, we're getting closer to a future where heart disease prediction is not just a research topic, but a reliable part of everyday healthcare, potentially saving countless lives.

6. Conclusion

In wrapping up this work, it's clear that machine learning can make a meaningful impact in predicting heart disease—potentially changing how we approach early diagnosis and treatment. Through experimenting with several models, it became obvious that advanced ensemble techniques like Random Forest and XGBoost don't just offer better accuracy, but also add a layer of reliability to predictions. This is crucial in a real-world setting where every correct diagnosis can mean a chance at early intervention and better outcomes for patients. At the same time, this project highlighted some real limits: working with a relatively small and possibly unbalanced dataset means that there's still work to do before we can confidently say these models will perform just as well in every clinic or across all patient populations.

References

World Health Organization. Cardiovascular diseases (CVDs) factsheet. **https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)** (2025)

Biswas N, et al. Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. Wiley, 2023. **https://onlinelibrary.wiley.com/doi/10.1155/2023/6864343**

World Health Organization. Uncontrolled high blood pressure puts over a billion people at risk. **https://www.who.int/news/item/23-09-2025-uncontrolled-high-blood-pressure-puts-over-a-billion-people-at-risk** (2025)

Chong B, et al. Global burden of cardiovascular diseases: projections from 2025 to 2050. European Journal of Preventive Cardiology, 2025. **https://pubmed.ncbi.nlm.nih.gov/39270739/**

Ahmad AA, et al. Prediction of Heart Disease Based on Machine Learning. PMC, 2023. **https://pmc.ncbi.nlm.nih.gov/articles/PMC10378171/**

El-Sofany H, et al. A proposed technique for predicting heart disease using machine learning. Nature Scientific Reports, 2024. **https://www.nature.com/articles/s41598-024-74656-2**

Kumar R, et al. A comprehensive review of machine learning for heart disease prediction. Frontiers in Artificial Intelligence, 2025. **https://www.frontiersin.org/journals/**artificial-**intelligence/articles/10.3389/frai.2025.1583459/full**

UCL Machine Learning Repository – Heart Disease Dataset. **https://archive.ics.uci.edu/dataset/45/heart+disease**

Detrano, R., Janosi, A., Steinbrun, W., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology. **https://pubmed.ncbi.nlm.nih.gov/2756873/**

Rahman, H., et al. (2024). Deep learning in cardiovascular disease detection: Opportunities and pitfalls. IEEE Access. **https://www.nature.com/articles/s41598-025-09594-8**

Framingham Heart Study, National Heart, Lung, and Blood Institute. **https://www.framinghamheartstudy.org/fhs-about/**

World Heart Report 2025.**https://world-heart-federation.org/wp-content/uploads/World_Heart_Report_2025_Online-Version.pdf**

An active learning machine technique-based prediction of heart diseases: **https://www.nature.com/articles/s41598-023-40717-1**

A proposed technique for predicting heart disease using machine learning algorithms: **https://www.nature.com/articles/s41598-024-74656-2**

Heart Disease Detection: A Comprehensive Analysis of Machine Learning Techniques

:**https://www.sciopen.com/article/10.26599/NBE.2024.9290087**