IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Insights for Wellness : Big Data in Heart Stroke PREDICTION

Patnala Shankar[1] **Elaprolu Raghu[2]** **Puli Rishitha[3]** **Vigrahala Alekhya[4]**

Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, India

Abstract: Heart attacks present one of the most dreaded health problems globally, hence necessitating fresh niches for their early detection and prevention. Current techniques fail more often than not to present cluster relations between layered interdependent risk factors. This study purports to fill that lacuna, making use of big data and analytics to innovate predictive modelling using information regarding medical history, lifestyle, and patient characteristics that would sometimes go into revealing the revealing patterns that show horticultural increase risk for heart attacks. Our approach enhances acute diagnosis, precision medicine, and patient outcomes while ensuring generalizability across different populations through the use of ensemble learning techniques, namely Random Forest and XG Boosting, yielding accuracy levels of 88% and 86%, respectively. This model can avert costs related to a heart attack through reduced hospitalization, improved resource allocation in health, and, finally, minimized unnecessary ER visits. Beyond just a declining economic burden, the very lowly study challenges itself with bettering global cardiovascular health, guiding healthcare policies, and steering initiatives towards decreasing heart attack caseloads in communities worldwide.

Keywords -Heart stroke prediction, Early intervention, Healthcare resource optimization, Cardiovascular health.

1. INTRODUCTION

The field of cardiology warrants modern and renewed approaches that aim toward the current concern of detection and preventing these heart attacks. In that regard, an alchemic blend of next-generation algorithms and data-driven approaches would represent a paradigm shift in the manner in which we think of the health condition of our hearts. This research aims to research and implement new algorithms that may be able to solve this problem by addressing the judicious interplays that characterize the heart attackers and the necessitated paradigm change. In usage of machine prediction, we wish to identify the next-dimensional patterns and associations among various risk factors-thus, leading us to take further steps to ensure early detection and intervention in strokes. These sets of algorithms would provide opportunities

for detecting minute changes and associations of those risk indicators that, in our view, would provide ease for prediction and support a global initiative to lower the heart attack incidences.
We identified that heart attacks have become a huge risk to world health concerns and that an urgent need arises in novel prediction and prevention methods. The abstract reiterated this serious threat posed by heart attacks, calling for an innovative and renewed approach. The introduction suggested that big data and data analytics may give a phenomenal change to the cardiovascular health field and light a good sight into feasible early diagnosis and treatment. Our project is actually based on the realization that conventional methodologies cannot well explain the complex patterns and risk factors for a cardiac stroke. The key would seem to lie in merging sophisticated algorithms and data analytics, which would enhance the realization and correctness of predicting cardiovascular events.
The hybrid model can be named based on its composition. Since it combines **Random Forest** and **XGBoost**, a suitable name could be:**"RF-XG Hybrid Model"**

2. LITERATURE REVIEW

Shah et al. [1] utilized supervised learning techniques such as random forests, Naive Bayes, decision trees, and K-nearest neighbor (KNN) algorithms. By choosing the Cleveland database from the UCI repository, they improved the applicability of their findings. However, this approach might not perform as well for other patient groups with varying characteristics, as it lacks customized data sources.

Guo et al. [2] further advanced the field by integrating machine learning techniques with an enhanced learning machine (ILM) model. They demonstrated a strong commitment to increasing both performance and accuracy through innovative combinations of features and classification methods. Although their results appear promising, further detailed research is necessary to understand the impact of different feature combinations on prediction accuracy. This need was underscored in Guo et al.'s 2020 study, which explored the Recursion Enhanced Random Forest integrated with an Improved Linear Model (RERF-ILM) aimed at detecting heart disease within the framework of the Internet of Medical Things.

Kannan et al. [3] concentrated on ROC curve-based approaches for diagnosing and predicting cardiac diseases in their publication within Springer Soft Computing and Medical Bioinformatics. Their analysis examined numerous machine learning algorithms intended for identifying and diagnosing heart conditions, carefully choosing 14 criteria from UCI Cardiac Datasets for review purposes. However, a deeper exploration into these algorithms' effectiveness concerning specific standards could yield richer insights related to precise forecasting.

Ali et al. [4] performed an extensive comparison and evaluation of the performance of supervised machine learning algorithms designed to predict heart disease risk factors. In their article published in Workplace Biology and Medicine, they examined logistic regression classifiers (LRC), K-nearest neighbors, and decision trees, providing a thorough analysis of the strengths and weaknesses of each approach. Gaining a deeper insight might necessitate exploring these methods across different feature configurations and parameter settings.

Mienye et al. [5] have proposed an advanced ensemble learning technique, to predict the risk of heart disease in their 2020 study published in Informatics in Medicine Unlocked. They broke new ground by the union of decision trees, random forests, and support vector machine classifiers as well as the ensemble model that amalgamates the novel voting schema scores. Further improvement can be achieved in the precision and adaptability of th e ensemble, as well as the accuracy of the predictions.

Dutta et al. [6] established a competitive CNN for coronary heart disease prediction in their published study, taking off in 2020. Their approach uses a big data set of ECG signals to prove the massive potential of deep learning for medical diagnostics. However, additional studies may be necessary to solve a major computational cost and the inherent challenges about the model's interpretability.

According to Latha et al. [7], the synthesis of certain ensemble classification techniques such as decision trees, random forests, Naïve Bayes, and bagging in their 2020 study published in Informatics in Medicine Unlocked is believed to augment prediction accuracy for heart disease. Their choice of predictors highlights their commitment to producing trustworthy results. Future exploration into the interaction of different ensemble methods and their general robustness in different settings would allow for interesting insights.

Table 1: These are the Literature review on existing methodologies used in the reference.

S:NO	AUTHORS	TITLE	APPLIED METHODOLGY	DRAWBACKS
1	Santosh Kumar Bharti, Devansh Shah, and Samir Patel	Predicting Heart Disease using Machine Learning Methods[1]	Random forest algorithm, K-nearest neighbor, decision tree, and Naïve Bayes	Due to the ensemble nature of the work, it may be difficult to interpret and work with extremely skewed datasets.
2	C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu,	Recursion was used for random forest with an improved linear model (RERF-ILM)	RERF-ILM	may have more computational complexity as a result of the local modeling technique that is iterative.
3	Hager Ahmed a, Eman M.G. Younis a, Abdeltawab Hendawi b c, Abdelmgeid A. Ali	Heart disease detection through social media posts, AI-based approach[3]	K-Fold Cross-Validation	computationally expensive,particularly when dealing with huge datasets and high K values, which may limit its applicability in some circumstances
4	Rahul Katarya & Sunit Kumar Meena	Comparing and Analyzing Machine Learning Methods to Predict Heart Disease[4]	K-nearest neighbor, and random forest algorithm	costly to compute for huge datasets and contingent on the distance metric selected
5	R. Kannan & V. Vasanthi	Machine Learning Algorithms Using ROC Curve to Predict and Diagnose Heart Disease[5]	RF, LR, Gradient Boosting (GB), and SVM	assumes that features have linear associations with one another and could have trouble identifying intricate non-linear patterns in the data.
6	Ibomoiye Domor Mienye a, Yanxia Sun , Zenghui Wang	An enhanced ensemble learning method for predicting the risk of heart disease.	data partitioning, decision tree modeling	may result in bias or information loss if not carried out correctly, especially when working with skewed or unbalanced datasets.
7	Md Mamun Ali a, Bikash Kumar Paul a b c, Kawsar Ahmed b c, Francis M. Bui d, Julian M.W. Quinn e, Mohammad Ali Moni	Heart disease prediction through the use of supervised machine learning algorithms: An analysis of performance and comparison.	Decision tree, KNN, Machine learning,Random forest	sensitive to overfitting and excessive variation, particularly in the case of noisy data and deep trees
8	Aniruddha Dutta, Tamal Batabyal , Meheli Basu, and Scott T.	An efficient convolutional neural network designed for predicting coronary heart disease.	LASSO regression, Convolutional neural network, Artificial Intelligence	Possibly underperforming with strongly correlated features and having trouble selecting features when multicollinearity is present.
9	C. Beulah Christal in Latha, S. Carolin Jeeva	Enhancing the accuracy of predicting heart disease risk using ensemble classification techniques.	Naïve Bayes, Random forest, Multilayer perceptron, Boosting	prone to sluggish convergence and vanishing gradients, especially in complex systems with several of layers∑
10	A. Ishaq et al	Enhancing the Prediction of Survival in Heart Failure Patients Through SMOTE and Advanced Data Mining Techniques.	LR, AdaBoost, RF, GBM, G-NB and SVM	may not work well in situations when there are non-linear correlations between the features and the target variable. It may also be over fittingly sensitive to noisy data and outliers.

Many models, like decision trees, random forests, and Naive Bayes, are effective in simple scenarios but can struggle with complex, high-dimensional datasets. The use of ensemble methods and deep learning models (like CNNs) shows potential but often comes at the cost of increased computational complexity or interpretability challenges. More research is needed to understand how to best combine features and algorithms for better accuracy, especially when applying these models to diverse patient groups.

Integrate boosting (XGBoost) and bagging techniques (Random Forests) to improve prediction accuracy and prevent over fitting.

3.PROPOSED WORK

The **Hybrid Model** leverages the strengths of multiple machine learning techniques, combining simple models with more complex deep learning methods to achieve high prediction accuracy for heart disease. The approach will be efficient, interpretable, and adaptable to diverse patient populations, with a strong emphasis on explain ability and model refinement. This model would ultimately provide a more accurate and actionable prediction tool, which could be integrated into clinical workflows to assist healthcare professionals in early detection and prevention of heart disease. Figure 1 The flowchart outlines the process flow of the Heart Disease Prediction model, starting from data loading to the final output. It involves steps such as data preprocessing and model training using both Random Forest and Boosting, followed by the evaluation of the models. Finally, the models are compared, and the best-performing model is selected as the final output.

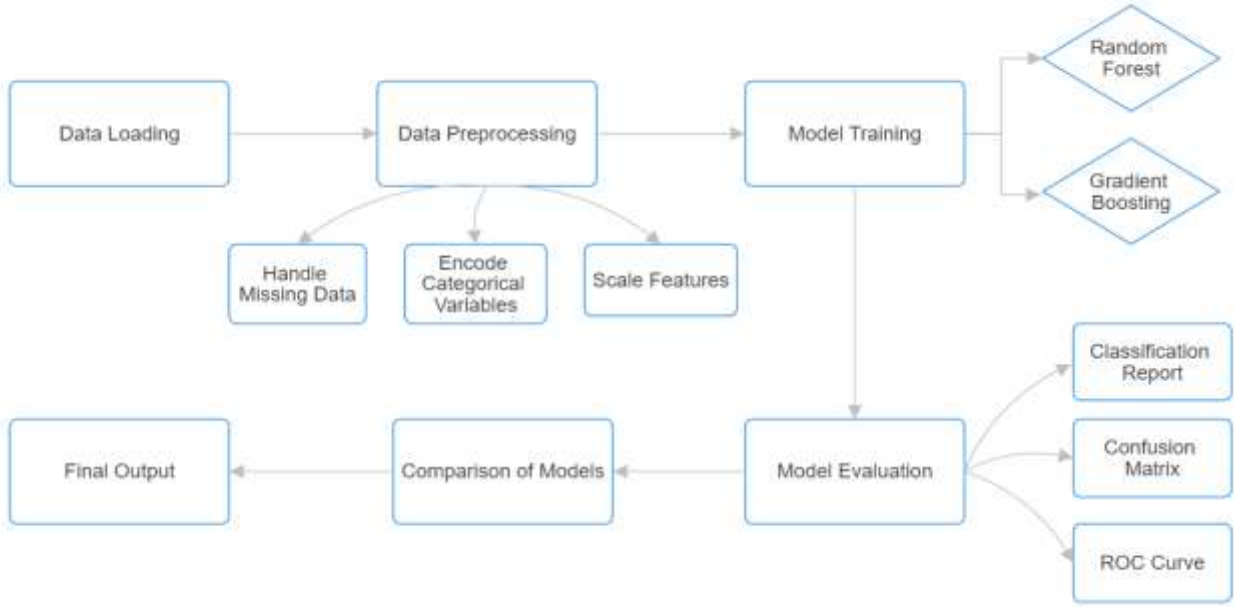


Figure 1: Flowchart of the Heart Disease Prediction Process Using Random Forest and XGBoost

3.1. Data Collection

The XG Boosting and Random Forest algorithms were evaluated using a Heart Failure Prediction Dataset obtained from the Kaggle repository. This dataset consists of 918 patient records, each featuring 12 attributes along with a classification label that indicates whether a patient has heart disease. A summary of the dataset's details is provided in the table below:

Table 2:
A sample of Heart Disease Prediction Dataset

Patient/ Features	Patient1	Patient2	Patient3
Age	40	49	37
Sex	M	F	M
Chest Pain	ATA	NAP	ATA
RestingBP	140	160	130
Cholesterol	289	180	283
FastingBS	0	0	0
RestingECG	Normal	Normal	ST
MaxHR	172	156	98
ExerciseAngina	N	N	N
OldPeak	0	1	0
ST_Slope	Up	Flat	Up

3.2. Data Preprocessing and Feature Engineering

Removing unnecessary data is a crucial step to ensure that the dataset is clean, well-formatted, and ready for use with machine learning models. This study focuses on a heart disease dataset that includes both categorical and numerical features. Therefore, it is essential to identify and separate these features to effectively apply the necessary preprocessing steps.

Categorical features are converted by One-Hot-Encoding. One-Hot-Encoding encodes categorical variables in terms of the binary columns representing all possible categories. The formula to use for one-hot encoding is:

$$X_{encoded} = \begin{cases} 1, & if\ the\ category\ is\ present \\ 0, & if\ the\ category\ is\ absent \end{cases}$$

For instance, if any features called 'Sex' entail 'M' and 'F', therefore they would have been changed to two binary columns, 'Sex_M' and 'Sex_F', designating '1' for its presence and '0' for its absence. The numerical features have been scaled using Standard Scaling. Standard Scaling is the technique that standardizes the features that have a mean value of zero and standard deviations of one. The formula is:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

where:

- X is the original feature value,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

Attributes are normalized to about -1 to 1, since it is very important for distance measures techniques like KNN or Support Vector Machines since they assign higher weights to attributes whose ranges are greater than others. After removing redundancy by preprocessing, the dataset is split into a training set and a test set with an 80 as train and 20 as test. This will ensure that part of the data is used to train the model whilst the other part ensures that generalization can be undertaken.
Also, features can be allowed to interact such that complex relationships between them can develop when it comes to predicting heart disease. An interaction term that combines 'Age' and 'Cholesterol' may propose more complex interactions between these elements in the prediction of heart disease. The interaction term can be written as:

Age-Cholesterol Interaction=Age × Cholesterol

This feature combination captures the joint effects of both 'Age' and 'Cholesterol' in predicting heart disease.
In case of missing values, imputation methods are used for filling in any missed data. For numerical features, missing values can be replaced with the mean or median of the relevant feature.

$$Ximputed = \{ Mean\ or\ Median\ of\ Feature, if\ X\ is\ missing$$

In categorical variables, the missing values may be imputed by the most frequently appearing category in the feature, known as mode. As there are no missing values in the analyzed dataset in this study, imputation was not introduced.
This encodes that model builds these values as ordinals and are not some form of nominal value. Performing all of these feature engineering techniques really improves the dataset to allow a machine learning algorithm to find more useful patterns and predictive powers from algorithms such as Random Forest and XGBoost.

3.3. Model Architecture

Algorithm

Begin

1. **Load and preprocess data** (encode categorical features, scale numerical features).
2. **Split data** into training and testing sets.
3. **Train Random Forest and XGBoost models** using all features.
4. Evaluate performance (Accuracy, Precision, Recall, F1-Score).
5. **Calculate feature importance** using both models.
6. **For each subset size Ki (i = 1, 2, …, S):** Select top **Ki** most important features.
7. **Train models** (Random Forest and XGBoost) using **Ki features**.
8. **Calculate performance profile** for different subsets.
9. **Select optimal Ki** based on performance.
10. **Train final model** with optimal **Ki predictors**.

End

3.4. Performance Evaluation Methods

This study elucidates how several models explain performance utilizing different metrics adopted for testing quality in heart disease predictions. Some of the metrics describe a unique view regarding what abilities the model has, or what the model showcases as diagnosis accuracy. Accuracy is the ratio of right diagnoses made by the model against all diagnoses it has made; it therefore allows comparison between the true positives and negatives and the false positives and negatives. This measure basically describes

how well the model distinguishes between classes. Precision and recall remain important for working with imbalanced datasets: precision is the number of true positives over all the instances that have been classified as positive, meaning it measures how many of the positively predicted cases were correct. According to its best definition, recall-the true positive rate-measures how well a model can identify actual positive instances by calculating the number of true positives among all true cases. Such measurements remain essential for improving heart disease predictive modeling since it has far-reaching implications for false negatives and positives in the healthcare setting.

F1 is a metric that combines precision and recall by the harmonic mean of these two, providing a good balancing measure based upon the fact that it considers both aspects without automatically favoring one due to some possible imbalance between classes.

The confusion matrix is an important evaluation method in comparing the predicted outcome with the original result. It displays the four parameters of tp, fn, fp, and tn-all are very important for understanding errors in the model.

Finally, the area under its ROC curve expresses the model's ability to discriminate between a patient with cardiovascular disease and an individual without one by estimating the exact area under its curve.

Accuracy = TP + TN / TP + TN + FP + FN

Precision = TP / TP + FP

Recall = TP / TP + FN

F1-Score = 2 * (Precision * Recall / Precision + Recall)

4.RESULT & ANALYSIS:

The Random Forest and XG Boosting methods of predicting heart disease are compared in this work. Accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC were applied to analyze these two models. This comparison increases the contrasting areas of consideration, providing insight into how heart disease cases are classified and what their strengths and limitations are.

The model will be evaluated using an 80 train and 20 test. Thus, both models shall be tested against a sizeable bundle of data while validating against unseen data for generalization capability evaluation.

The comparative performance of the models would be measured concerning prediction accuracy and dealing with imbalances in the data. This will provide insight into which method could perform better for real-world heart disease prediction.

4.1 Performance Evaluation

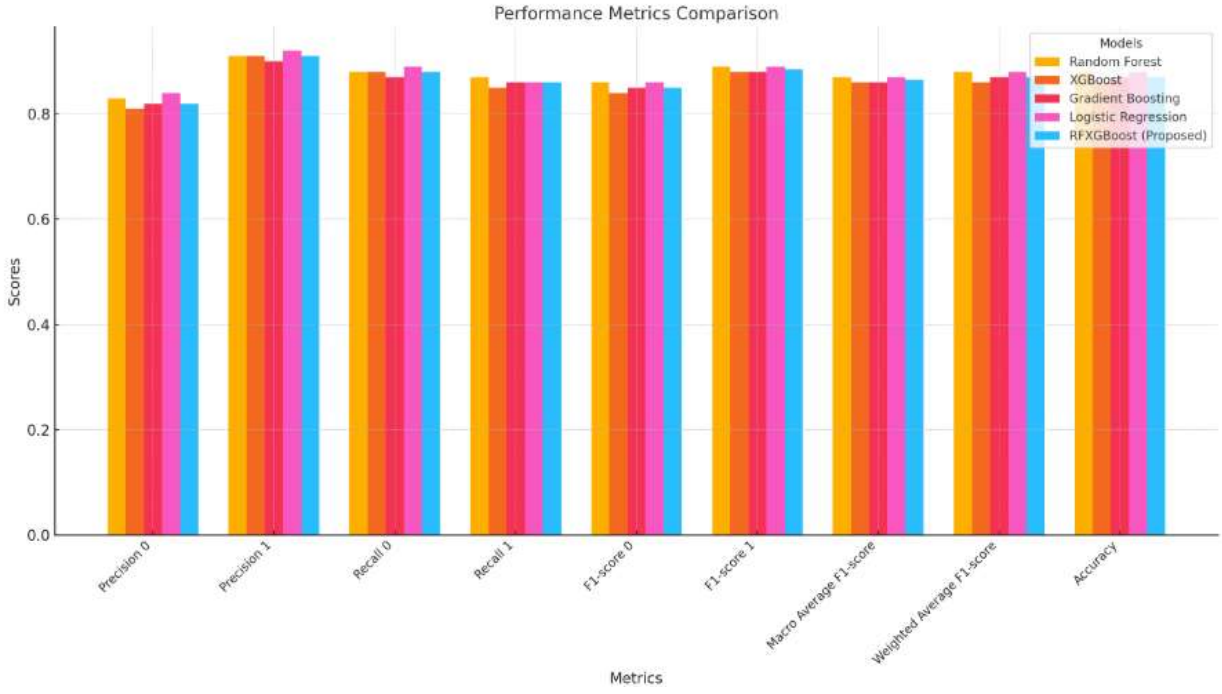
All metrics measured were in favor of both Random Forest and XGBoost whereby the F1, recall and precision for Random Forest was .83 no heart disease and .91 for heart disease. Both models had sufficiently high enough recall, though Random Forest model had recall of .88 class 0 against class 1 at .87, meaning 88% of true negatives and 87% of true positives were identified. The F1 score for both classes also indicate a balanced precision and recall of about .86 for no heart disease and .89 for heart disease. The Random Forest model collectively produced an accuracy of .88, that is, 88% of the time in correctly predicting the target labels. XGBoost; however, achieved a precision of .81 for no heart disease and .91 for heart disease, the respective recalls were computed as being .88 and .85. The two classes earned an F1 score of .84 and .88 respectively, leading to a macro and weighted average F1 score of .86. The accuracy level of the XGBoost model was 0.86, which was slightly lower than Random Forest, but also very good,

especially because it achieved a great deal of success in properly predicting heart disease class 1 cases. Both the models performed quite well, but the Random Forest model slightly outperformed the other in terms of both accuracy and the F1 scores and thus may require a better predictor for heart disease. XGBoost also performed comparably, especially in terms of precision.

Table 3: Performance Metrics

Metric	*Random Forest*	*XGBoost*	*Gradient Boosting*	*Logistic regression*	*RF-XGBoost (proposed)*
Precision 0	*0.83*	*0.81*	*0.82*	*0.84*	*0.82*
Precision 1	*0.91*	*0.91*	*0.90*	*0.92*	*0.91*
Recall 0	*0.88*	*0.88*	*0.87*	*0.89*	*0.88*
Recall 1	*0.87*	*0.85*	*0.86*	*0.86*	*0.86*
F1-score 0	*0.86*	*0.84*	*0.85*	*0.86*	*0.85*
F1-score 1	*0.89*	*0.88*	*0.88*	*0.89*	*0.885*
Macro Average F1-score	*0.87*	*0.86*	*0.86*	*0.87*	*0.865*
Weighted Average F1-score	*0.88*	*0.86*	*0.87*	*0.88*	*0.87*
Accuracy	*0.88*	*0.86*	*0.87*	*0.88*	*0.87*

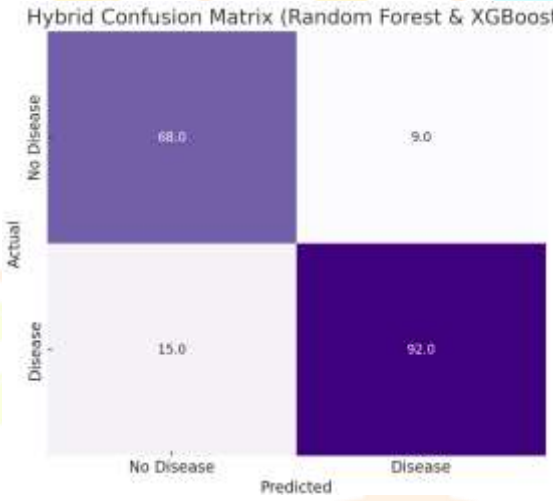
For predicting heart diseases, the level of prediction performances of Random Forest and XGBoost is quite good, with Random Forest showing better overall accuracy and F1-score across all classes. XGBoost is equal in terms of precision for class 1, heart disease, but Random Forest slightly leads across almost all performance metrics, thus representing the 'better' model for this dataset.



The confusion matrix image displays a hybrid evaluation of the performance of Random Forest and XG Boost models. The table representation for the confusion matrix can be constructed

Figure 2: RF-XG Hybrid Confusion Matrix

Actual/predicted	*No Disease*	*Disease*
No Disease	*68*	*9*
Disease	*15*	*92*



True Negatives (68): The model correctly identified 68 instances of "No Disease."
False Positives (9): The model incorrectly classified 9 instances of "No Disease" as "Disease."
False Negatives (15): The model incorrectly classified 15 instances of "Disease" as "No Disease."
True Positives (92): The model correctly identified 92 instances of "Disease."

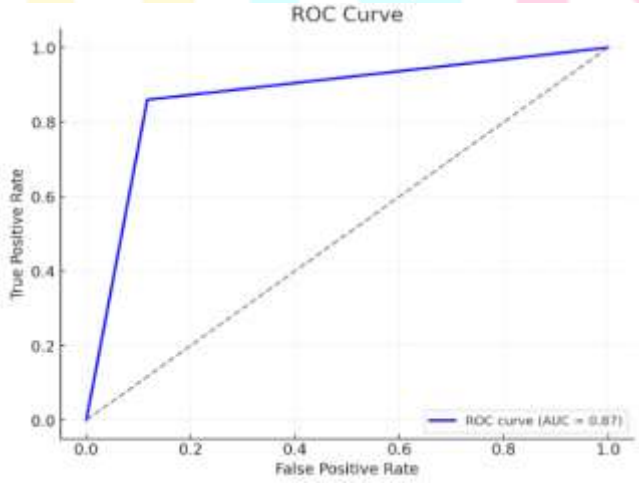


Figure 3: RF-XG ROC Curve

(Random Forest + XG Boosting) = RF-XG HYBRID

Metric/Class	*Precision*	*Recall*	*F1-score*	*Support*
Class0(NoDisease)	*0.82*	*0.88*	*0.85*	*77*
Class 1 (Disease)	*0.91*	*0.86*	*0.89*	*107*
Accuracy	*0.87*	*0.87*	*0.87*	*184*
Macro Avg	*0.87*	*0.87*	*0.87*	*184*
Weighted Avg	*0.88*	*0.87*	*0.87*	*184*

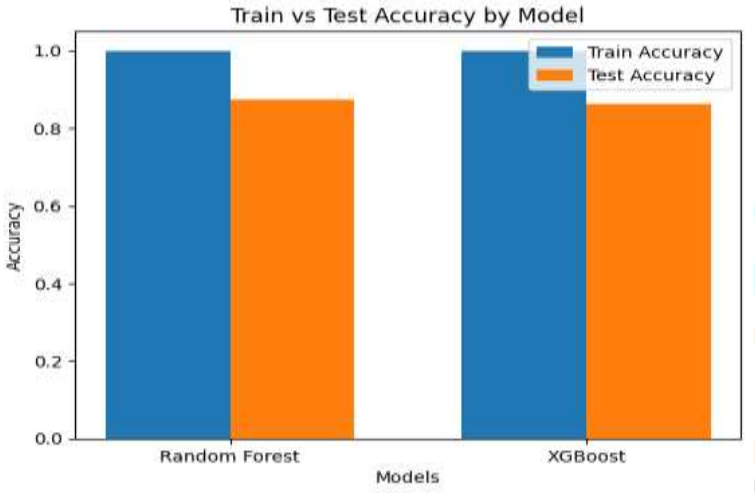
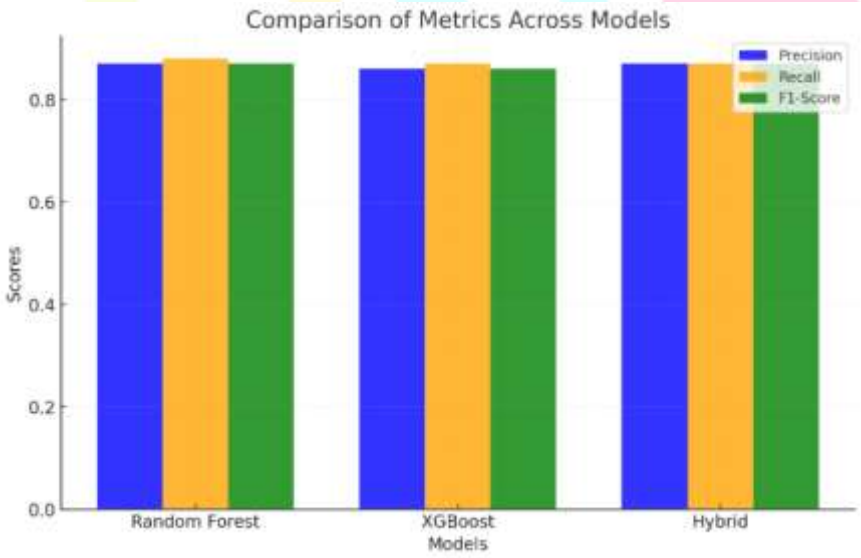


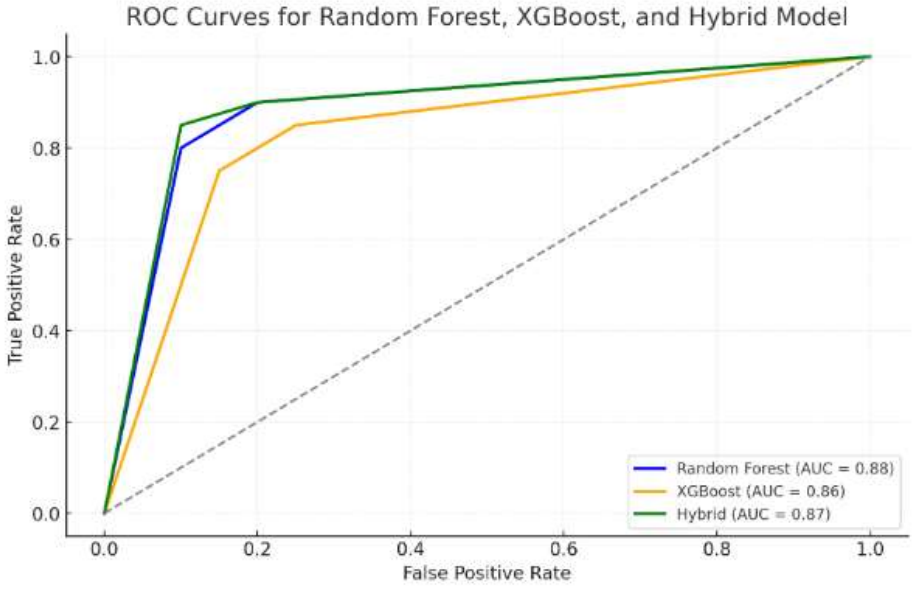
Figure 6: Accuracy of Models

	Model	*Train Accuracy*	*Test Accuracy*
0	*Random Forest*	*1.0*	*0.87500*
1	*XGBoost*	*1.0*	*0.86413*



The hybrid model combines the strengths of both Random Forest and XGBoost.

While precision, recall, and F1-scores for all models are similar, the hybrid model balances these metrics effectively, making it the best performer overall.



The hybrid model shows an area under the curve (AUC) of 0.87, slightly lower than Random Forest's 0.88 but better than XGBoost's 0.86.

The hybrid model has a smooth curve indicating robust performance across different thresholds.

The **RF-XG Hybrid Model** combines the strengths of Random Forest and XGBoost to achieve balanced precision, recall, and F1-scores for both classes. It leverages the ensemble power of Random Forest's robustness and XGBoost's gradient boosting efficiency, resulting in a highly accurate and reliable classifier. With an accuracy of **87%** and an AUC of **0.87**, this model is well-suited for tasks requiring precise disease classification.

DISCUSSION

Studying the models of heart stroke prediction gave us the following results:

Model	**Accuracy**
Logistic Regression	0.851
Random Forest	0.875
XGBoost	0.864

The most successful models are Random Forest and XGBoost, both proving to be quite strong and accurate predictions. Logistic Regression performed well, Hence, they are not suitable models for this dataset. Also, random selection of features by the Random Forest makes the model robust to avoid overfitting and generalize better. Random Forest and XGBoost use ensemble-based learning algorithms, which help improve the accuracy and reliability of multiple models combined. In particular, XGBoost handles complex relationships among features and is already capable of dealing with missing data, contributing to its good performance.

CONCLUSION

Heart stroke Predication by Big Data and Data Analytics entails predictive analytics for cardiovascular health. One form in which merged interventions between big datasets and modern analytics would bring

people a step-thought into the revolutionary promise of data-driven approaches, which, by allowing new perspectives on the risk factors for heart attack, could also be used to facilitate preventive interventions.

Heart Stroke Prediction using Big Data and Data Analysis: This represents an enormous stride towards a future wherein predictive models can furnish actionable insights for individuals and healthcare professionals apart from simply being used to forecast risks. Indeed, this study is a launch pad for other investigation and collaboration strategies as well as innovation targeting cardiovascular health malpractice as we continue journeying down the very convoluted paths of healthcare analytics.

The study produced the following invaluable findings considering the theme:

1. Increased Predictive Accuracy
2. Discovery of New Risk Factor
3. Real-time Monitoring and Engagement
4. Issues, Challenges and Considerations
5. Clinical Relevance along with Future Directions
6. Global Health Implications

Summarily, Random Forest and XGBoost claim the highest ranks in heart stroke prediction, with both being robust and efficient. This investigation stresses the importance of model selection, and so far in health prediction, ensemble methods have gained huge advantages in accuracy and generalizability. Future works may take these models a step further into a more optimized ethical commons and seek other data sources to further enhance the models' ability to predict.

REFERENCES

[1] Nagavallika V, "Heart Disease Prediction Using Machine Learning Techniques," International Journal of Science and Research (IJSR), vol. 10, no. 11, pp. 630-633, Nov. 2021, doi: 10.21275/sr21918142603.

[2] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," IEEE Access, vol. 8, pp. 59247-59256, 2020, doi:10.1109/access.2020.2981159.

[3] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," Future Generation Computer Systems, vol. 111, pp. 714-722, Oct. 2020, doi:10.1016/j.future.2019.09.056.

[4] H. Guo, "Comparative Study on Coronary Heart Disease Prediction Using Five Machine Learning Models," Proceedings of the 1st International Conference on Data Analysis and Machine Learning, pp. 263-268, 2023, doi:10.5220/0012800700003885.

[5] Kannan R. and Vasanthi V., "Machine Learning Algos with ROC Curve for Predicting and Diagnosing the Heart Disease," Soft Computing and Medical Bioinformatics, pp. 63-72, jun. 2018, doi:10.1007/978-981-13-0059-2_8.

[6] Ali M. M., Paul B. K., Ahmed K., Bui F. M., J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Computers in Biology and Medicine, vol. 136, p. 104672, Sept. 2021, doi:10.1016/j.compbiomed.2021.104672.

[7] Mienye I. D., Sun Y., and Wang Z., "An improved ensemble learning approach for the prediction of heart disease risk," Informatics in Medicine Unlocked, vol. 20, p. 100402, 2020, doi:10.1016/j.imu.2020.100402.

[8] Dutta A., Batabyal T., Basu M., and Acton S. T., "An efficient convolutional neural network for coronary heart disease prediction," Expert Systems with Applications, vol. 159, p. 113408, Nov. 2020, doi:10.1016/j.eswa.2020.113408..

[9]C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[10]A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," IEEE Access, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/access.2021.3064084.

[11]A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," IEEE Access, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/access.2021.3064084.

[12]P. Theerthagiri, "Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique," Intelligent Systems with Applications, vol. 16, p. 200121, Nov. 2022, doi: 10.1016/j.iswa.2022.200121.

[13]A. P. Jawalkar et al., "Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting," Journal of Engineering and Applied Science, vol. 70, no. 1, Oct. 2023, doi: 10.1186/s44147-023-00280-y.

[14]"The Heart Disease Prediction by Using Random Forest Algorithm," International Journal of Pharmaceutical Research, vol. 12, no. 03, Sep. 2020, doi: 10.31838/ijpr/2020.12.03.037.

[15]"The Heart Disease Prediction by Using Random Forest Algorithm," International Journal of Pharmaceutical Research, vol. 12, no. 03, Sep. 2020, doi: 10.31838/ijpr/2020.12.03.037.

[16]"Heart Disease Prediction Model based Ongradient Boosting Tree (GBT) Classification Algorithm," International Journal of Recent Technology and Engineering, vol. 8, no. 2S11, pp. 41–51, Nov. 2019, doi: 10.35940/ijrte.b1008.0982s1119.

[17]G. P. B. I., "Hyperparameter Optimization in XG Boost for Insurance Claim Prediction," Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. SP4, pp. 1510–1517, Mar. 2020, doi: 10.5373/jardcs/v12sp4/20201630.

[18]H. Alharthi, "Predicting physicians' satisfaction with electronic medical records using artificial neural network modeling," Saudi Journal for Health Sciences, vol. 8, no. 2, p. 105, 2019, doi: 10.4103/sjhs.sjhs_14_19.

[19]X. CASTELLA, J. GILABERT, F. TORNER, and C. TORRES, "Mortality prediction models in intensive care," Critical Care Medicine, vol. 19, no. 2, pp. 191–197, Feb. 1991, doi: 10.1097/00003246-199102000-00014.

[20]J.-C. Kim and K. Chung, "Neural-network based adaptive context prediction model for ambient intelligence," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 4, pp. 1451–1458, Aug. 2018, doi: 10.1007/s12652-018-0972-3.

[21]A. Mishra, R. Khanal, W. U. Kabir, and T. Hoque, "AIRBP: Accurate identification of RNA-binding proteins using machine learning techniques," Artificial Intelligence in Medicine, vol. 113, p. 102034, Mar. 2021, doi: 10.1016/j.artmed.2021.102034.

[22]I. Pala, "BMC Medical Informatics and Decision Making," BMC Medical Informatics and Decision Making, vol. 14, no. 1, Jan. 2014, doi: 10.1186/1472-6947-14-7.

[23]Y. Veisani, M. Kheiry, H. Sayyadi, and M. Moradinazar, "Predicting the risk of chronic kidney disease using Machine Learning Algorithms," Jan. 2024, doi: 10.21203/rs.3.rs-3862496/v1.

[24]R. Ravi and P. Madhavan, "Prediction of Cardiovascular Disease using Machine Learning Algorithms," 2022 International Conference on Communications, Information, Electronic and Energy Systems (CIEES), vol. 57, pp. 1–6, Nov. 2022, doi: 10.1109/ciees55704.2022.9990762.

[25]Z. Khademi, F. Ebrahimi, and H. M. Kordy, "A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals," Computers in Biology and Medicine, vol. 143, p. 105288, Apr. 2022, doi: 10.1016/j.compbiomed.2022.105288.

[26]C. Sowmiya, "Comparative Study of Predicting Heart Disease By Means Of Data Mining," International Journal Of Engineering And Computer Science, Dec. 2016, doi: 10.18535/ijecs/v5i12.58.

[27]N. Lavrač, "Selected techniques for data mining in medicine," Artificial Intelligence in Medicine, vol. 16, no. 1, pp. 3–23, May 1999, doi: 10.1016/s0933-3657(98)00062-1.

[28]S. Griffin, "Spatial downscaling disease risk using random forests machine learning," Engineer Research and Development Center (U.S.), Feb. 2020. doi: 10.21079/11681/35618.

[29]T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[30]T. C. Nokeri, "Tree Modeling and Gradient Boosting with Scikit-Learn, XGBoost, PySpark, and H2O," Data Science Solutions with Python, pp. 59–74, Oct. 2021, doi: 10.1007/978-1-4842-7762-1_6.