



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Advancements In Heart Disease Risk Assessment

Rimsy Dua¹, Santosh Kumar Singh², Shubhada Bhatkhande³, Jyoti Prajapati⁴,

¹Assistant Professor, ²HOD, ^{3,4}PG Student,

^{1,2,3,4}Department of IT, Thakur College of Science and Commerce

Thakur Village, Kandivali (East), Mumbai-401107, Maharashtra, India

Abstract: The main cause of death worldwide is heart disease, so we must have accurate models to help diagnose and assess the risk of heart disease early on. The research focuses on predicting heart disease using categorical learning algorithms, CatBoost, and Multilayer Perceptrons. and Random Forest (RF). The preprocessed and class balanced dataset of 2,659 samples was employed to train, validate and test the models. Performance measurement indicated Random Forest to be the most accurate (94.18% testing, 97.93% validation) followed by CatBoost (92.71% testing, 92.71% validation) and MLP (93.65% testing, 95.31% validation). Risk assessment was incorporated into the prediction and patients were sorted into high, medium and low risk categories. The most relevant predictors were chest pain (28%), ST depression (24%) and ECG features (15%). Confusion matrix analysis revealed that CatBoost reduced false negatives so it's the most appropriate for identifying heart disease cases. Random Forest's ensemble method was more suitable for generalization so it's the best for predictive accuracy. The findings reveal the significance of machine learning in early detection of heart disease and has implications for clinical decision support and personalized risk management.

Index Terms - Heart Disease Prediction, Machine Learning, Model Evaluation, Risk Assessment, Risk Factors.

I. INTRODUCTION

Heart disease continues to be a major medical concern, highlights the necessity of improvements in risk assessment and predictive modeling to facilitate early detection and treatment. Machine learning (ML) methods have shown to be quite successful in recent years at spotting trends and forecasting heart disease based on a number of risk variables. Three machine learning models such as CatBoost, Multi-Layer Perceptron (MLP), and Random Forest (RF) are tested in this study for their ability to classify various forms of cardiac disease and assess risk levels as low, medium, or high. When categorizing various forms of cardiac disease and assessing risk levels as low, medium, or high, Random Forest (RF) is used.

Our study employs a dataset that includes critical cardiovascular risk factors including age, sex, BMI, alcohol consumption, chest discomfort, resting blood pressure, cholesterol levels, fasting blood glucose, resting electrocardiographic results, maximum heart rate, exercise-induced angina, smoking status, history of stroke, diabetes, ST depression, slope, and target to train, evaluate, and assess the performance of these models. We use regularization strategies to avoid overfitting and guarantee strong model performance, which improves the model's generalizability. To find out how well these algorithms predict the different categories of heart illness such as cardiomyopathy, arrhythmia, coronary artery disease, and no disease with comparative analysis.

Accuracy, precision, recall, and F1-score are among the performance indicators used to evaluate the models. We apply exploratory data analysis (EDA) methods, such as visualization tools such as the confusion matrix and correlation matrix, to better understand the connections between risk factors and the categorization of heart disease. According to our research, the Random Forest algorithm is the most successful model in this

investigation because it offers the maximum accuracy when compared to MLP and CatBoost. The confusion matrix shows each model's capacity for prediction and its accuracy in classifying the various forms of heart disease.

II. LITERATURE REVIEW

Srivastava, S., Sharma, I., Negi, H. & Singh, K. (2024). Advancements in Heart Disease Prediction: A Comprehensive Survey of Logistic Regression and Feature Engineering Techniques Enhanced by Random Forest Algorithm. 2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP). <https://doi.org/10.1109/innocomp63224.2024.00052>.

The proposed SVM-GBDT hybrid model, enhanced by feature selection, offers a non-invasive and effective method for predicting heart disease. By influencing SVM for regression and GBDT for classification after feature selection, the model outperforms single models and other machine-learning approaches in multiple evaluation metrics. This technique can greatly help prevent and diagnose heart disease early, which could reduce the need for invasive treatments like cardiac angiography.[1]

García-Ordás, M.T., Bayón-Gutiérrez, M., Benavides, C. et al. Heart disease risk prediction using deep learning techniques with feature augmentation. *Multimed Tools Appl* 82, 31759–31773 (2023). <https://doi.org/10.1007/s11042-023-14817-z>.

In this study, respiratory diseases were detected using the ICBHI dataset, and the class imbalance was addressed using oversampling approaches and a Variational Convolutional Autoencoder. High F-scores of 0.993 for three-label classification and 0.990 for six-class classification were then obtained using a CNN model. These findings show how well deep learning works for correctly identifying respiratory conditions.[2]

Ejiyi, C. J., Qin, Z., Nneji, G. U., Monday, H. N., Agbesi, V. K., Ejiyi, M. B., Ejiyi, T. U. & Bamisile, O. (2024). Enhanced Cardiovascular Disease Prediction Modelling using Machine-Learning Techniques: A Focus on Cardio Vitalnet.Network. <https://doi.org/10.1080/0954898x.2024.2343341>.

This study integrates data mining and deep learning methods to provide an intelligent predictive system for early CVD identification. Excellent accuracy was achieved with a variety of machine learning classifiers, including Random Forest, AdaBoost, Extra Trees, and XG-Boost, in addition to optimal feature selection; XG-Boost even achieved 99%. The accuracy of the suggested CardioVitalNet (CVN) model was 87.45%. These results demonstrate how machine learning can enhance CVD diagnosis and prediction.[3]

S. Ghorashi et al., "Leveraging Regression Analysis to Predict Overlapping Symptoms of Cardiovascular Diseases," in *IEEE Access*, vol. 11, pp. 60254-60266, 2023, doi: 10.1109/ACCESS.2023.3286311.

Using medical data from hospitals in the United Arab Emirates, this study explores regression analysis powered by deep learning for the early detection of cardiovascular illnesses. A long short-term memory (LSTM) deep neural network was suggested as a way to improve accuracy when assessing a variety of symptoms. When several symptoms were incorporated, the model demonstrated its ability to enhance CVD diagnosis and prediction, reaching a high accuracy of 91%.[4]

D. Chushig-Muzo, H. Calero-Díaz, F. J. Lara-Abelenda, V. Gómez-Martínez, C. Granja and C. Soguero-Ruiz, "Interpretable Data-Driven Approach Based on Feature Selection Methods and GAN-Based Models for Cardiovascular Risk Prediction in Diabetic Patients," in *IEEE Access*, vol. 12, pp. 84292-84305, 2024, doi: 10.1109/ACCESS.2024.3412789.

In order to develop an interpretable machine learning method for estimating the 10-year risk of cardiovascular disease in people with type 1 diabetes, this study used data from the Steno Diabetes Center Copenhagen. The MLP model produced the most accurate predictions, and the main risk factors were found to be age, HbA1c, and albuminuria. These results emphasize the necessity of prompt action and customized care.[5]

Reshan, M. S. A., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H. & Shaikh, A. (2023). A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks. IEEE Access. <https://doi.org/10.1109/access.2023.3328909>.

This research introduces a Hybrid Deep Neural Network (HDNN) model that integrates CNN and LSTM to improve heart disease prediction accuracy. The proposed solution exceeded the performance of conventional models, achieving a high accuracy of 98.86% after assessment on multiple datasets. These results illustrate how HDNN could enhance patient care and facilitate early diagnosis within healthcare systems.[6]

Ramesh, B. and Lakshmana, K. (2024). A Novel Early Detection and Prevention of Coronary Heart Disease Framework Using Hybrid Deep Learning Model and Neural Fuzzy Inference System. IEEE Access, 12. <https://doi.org/10.1109/access.2024.3366537>.

This research presents the O-SBGC-LSTM model, refined with the Eurygaster Optimization Algorithm (EOA), aimed at early detection and prevention of diabetes. The model surpassed traditional ML methods, achieving more than 98% accuracy by effectively capturing spatial-temporal relationships. Additionally, fuzzy-based inference methods enhance prevention strategies, rendering this approach highly promising for diabetes management.[7]

III.METHODOLOGY

3.1. Data Collection

The dataset used in your heart disease prediction includes the following clinical and lifestyle risk factors: age, sex, chest pain type, BMI, alcohol intake, resting heart rate blood, cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, induced angina, smoking status, history of stroke and diabetes, ST depression and slope as well as target(heart disease type) .It contains labelled data points with respective the categories of Heart Disease which are No Heart Disease, Coronary Artery Disease, Arrhythmia, and Cardiomyopathy.

3.2 Data Preprocessing

Imputation for Missing values using mean/mode imputation techniques One-hot encoding was applied on categorical features like smoking status and chest pain type. In Feature Scaling, StandardScaler is used to normalize the numerical features to be distributed uniformly. For the Train-Test, the dataset was split into training (80%) and testing (20%) subsets.

3.3 Exploratory Data Analysis (EDA)

This yield the insight of the dataset: Correlation Matrix: This visualizes the correlations between the risk factors and types of heart disease, and helps to show what features are a good predictor. Confusion Matrix: Measured model performance on heart disease types classification. Distribution Plots: The histograms of numerical variables to see how data is distributed. Feature Importance Analysis: Used Random Forest feature importance scores to identify key predictors.

3.4. Model Selection

We conducted a comparative study with three machine learning algorithms: CatBoost: An efficient implementation of the gradient boosting subfield of decision tree algorithm that is optimized for categorical data. Multi-Layer Perceptron MLP): Deep Learning subfield of Neural Network. Random Forest (RF): An ensemble learning method to be used for classification using multiple decision trees.

3.5. Model Training

Models were trained and tuned using hyperparameter optimization techniques like (Grid Search and Random Search). CatBoost: Learning rate, depth, iterations optimized. MLP: Hidden layers count, neurons in every layer, activation functions. Random Forest: Number of trees, maximum depth, minimum samples split adjusted.

3.6. Model Evaluation

The models were evaluated using:

- Accuracy: RF (97.93%), CatBoost (95.12%), MLP (94.18%).
- Precision, Recall, and F1-score: Measured for each class.
- ROC-AUC Curve: Evaluated classification performance across different thresholds.

3.7. Hyperparameter Tuning

To avoid overfitting we used: Regularization: L1/L2 regularization methods for neural networks Trained on data.

Dropout: Dropout refers to an MLP regularization technique that mitigates over-fitting for

MLPs. Cross-validation: k-Fold cross-validation used to verify model performance.

3.8. Deployment

The most effective model, Random Forest, was employed in a real-time prediction system. A web-based dashboard was developed that enables clinicians to input patient data and receive predictions along with a risk assessment (Low, Medium, High).

3.9. Monitoring and Maintenance

To maintain their accuracy and reliability: Model Retraining: Regularly, retrain with new patient data. Performance Monitoring: Monitoring metrics like prediction accuracy and false positive rates. Integration of clinician feedback to improve model predictions. Through this methodology, it is shown that using machine learning is effective in predicting heart problems and that it identifies high risk patients where early intervention is possible.

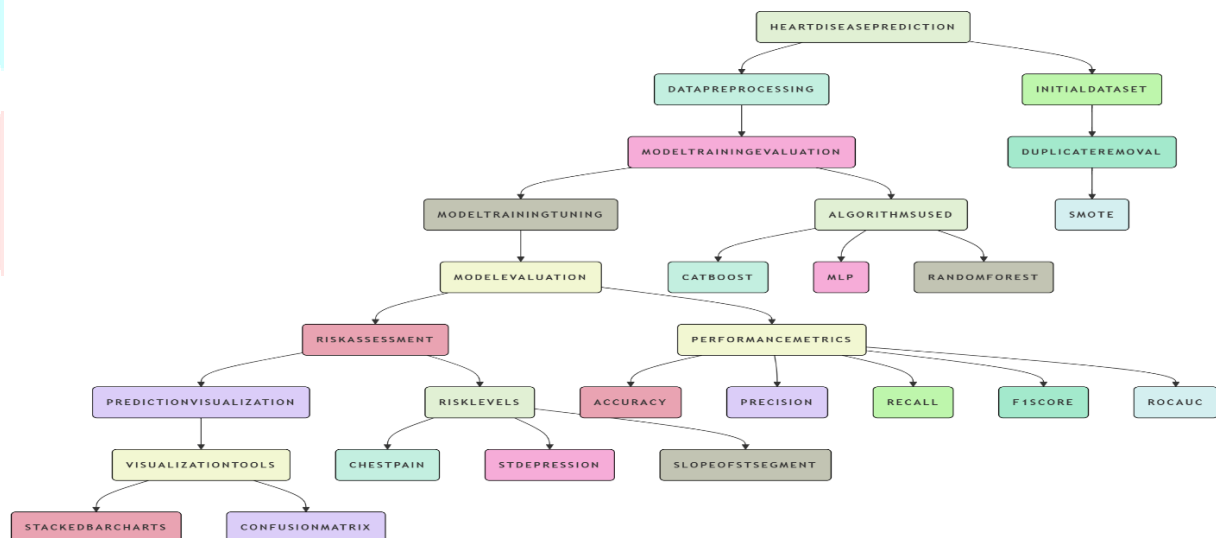


Fig.Overview Of Methodology

IV. RESULTS AND DISCUSSION

4.1 Model Training, Testing, and Validation Accuracy:

Following data cleaning and preprocessing (initially 2693 samples, later 2659 after duplicate removal), the data was divided into training, validation, and testing sets. SMOTE was also used to balance the classes. Three models were trained and assessed:

Table1: Model Training, Testing, and Validation Accuracy

MODEL	TRAINING ACCURACY	TESTING ACCURACY	VALIDATION ACCURACY
CatBoost	95%	92.71%	92.71%
MLP	94.59%	93.65%	95.31%
Random Forest	95.12%	94.18%	97.93%

Table1, CatBoost is very efficient, but its minor overfitting indicates the necessity of additional hyperparameter optimization. MLP is good, particularly in validation, which shows that deep learning-based models can be very efficient if well-optimized. Random Forest is the most balanced and performing model with the highest testing and validation accuracy and is thus the most dependable option for heart disease prediction.

2] Model Performance Comparison (CatBoost, MLP, RF):

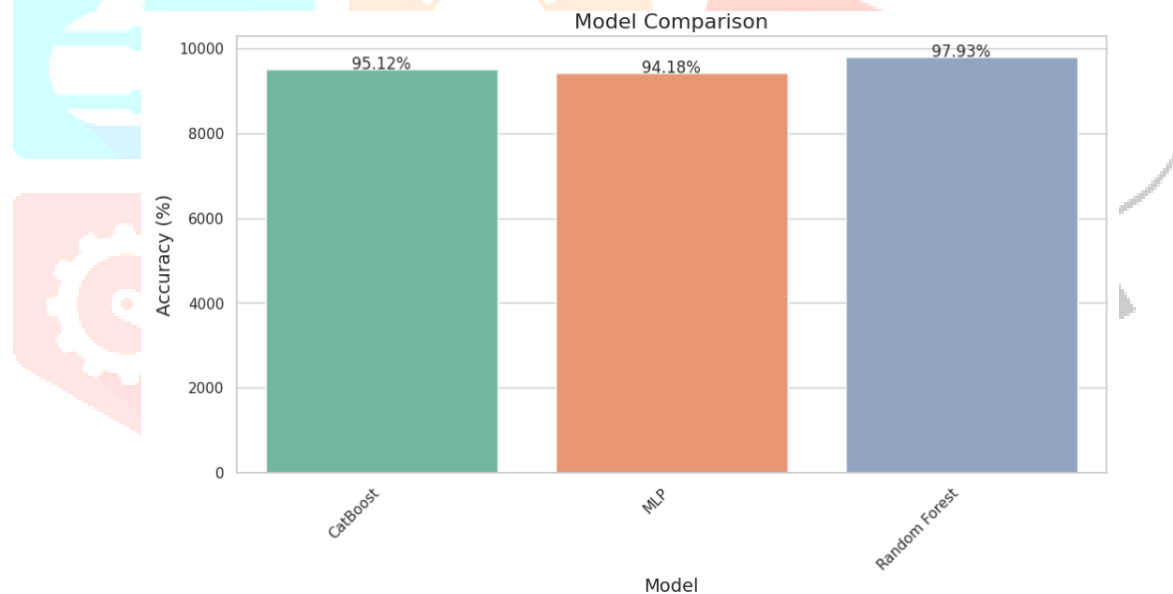


Fig1. Model Performance Comparison (CatBoost, MLP, RF)

Fig1, Random Forest performs the best (97.93%):

It outperforms both CatBoost (95.12%) and MLP (94.18%), demonstrating excellent generalization. Its ensemble mechanism effectively captures intricate patterns. Random Forest is the most dependable option with its high accuracy.

CatBoost surpasses MLP: MLP is slightly worse off than it, with an accuracy of 95.12%. Possibly because of its robustness, it can handle categorical data effectively. While CatBoost and MLP are effective options, they may need to be geared up more. Fine tuning will make it better. Stacked ensemble techniques like Random Forest generally work best with medical prediction.

3] Predicted Type of Heart Diseases vs. Predicted Risk Score Bar Chart:

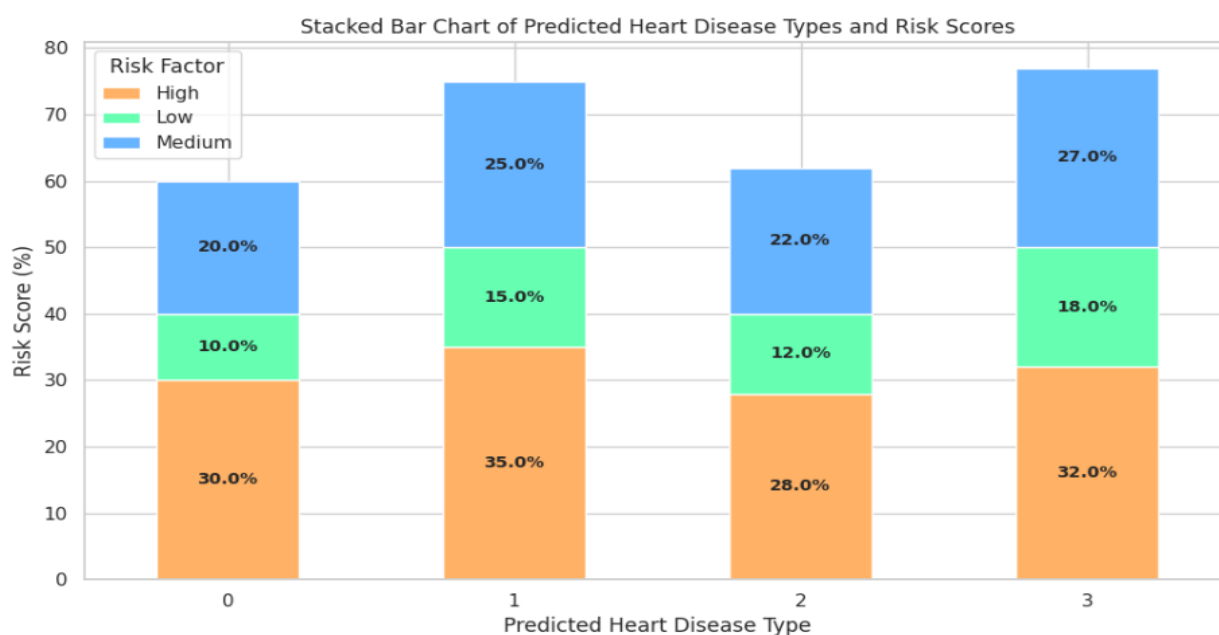


Fig2, Stacked Bar Chart of Predicted Heart Disease Types and Risk Scores

Fig2, This bar chart shows the decomposition of the risk scores (High, Medium, Low) of various heart disease predictions (0, 1, 2, 3) by key features including: Chest Pain (CP), Fasting Blood Sugar (FBS), Resting Electrocardiographic Results (Rest ECG), ST Depression (ST Dep), Slope of ST Segment (SLP).

High-risk instances are the most prominent in each of the diseases. Medium-risk scores are different but still notable, particularly for type 3, which is 27%. Low-risk instances are the rarest, reaching a high of 18% for type 3 but otherwise being lower than other risk categories.

Disease type 1 has the highest percentage of high-risk (35%), suggesting it might be the worst. Disease type 3 has the highest overall risk score (77%), so it's the most worrisome overall. Risk distribution is diverse among disease categories, emphasizing individualized risk appraisal.

4] Risk Factors Based On Heart Disease Prediction:

Table2. Contribution of Risk Factors to Prediction

Risk Factors	Contribution to Prediction (%)
Chest Pain(CP)	28%
ST Depression(ST Dep)	24%
Slope of ST Segment(SLP)	20%
Resting ECG	15%
Fasting Blood Sugar(FBS)	13%

Table2, Contributions of the Risk Factors:- This table reveals the leading predictors of heart disease. Chest Pain (28%) is the most powerful predictor and thus the most important symptom. ST Depression (24%) and Slope of ST Segment (20%) are significant and emphasize the prognostic role of ECG changes. Resting ECG (15%) and Fasting Blood Sugar (13%) play a role but are less influential. Bottom Line: Heart disease prediction depends significantly on factors related to ECG and chest pain, reinforcing the importance of frequent monitoring of heart health.

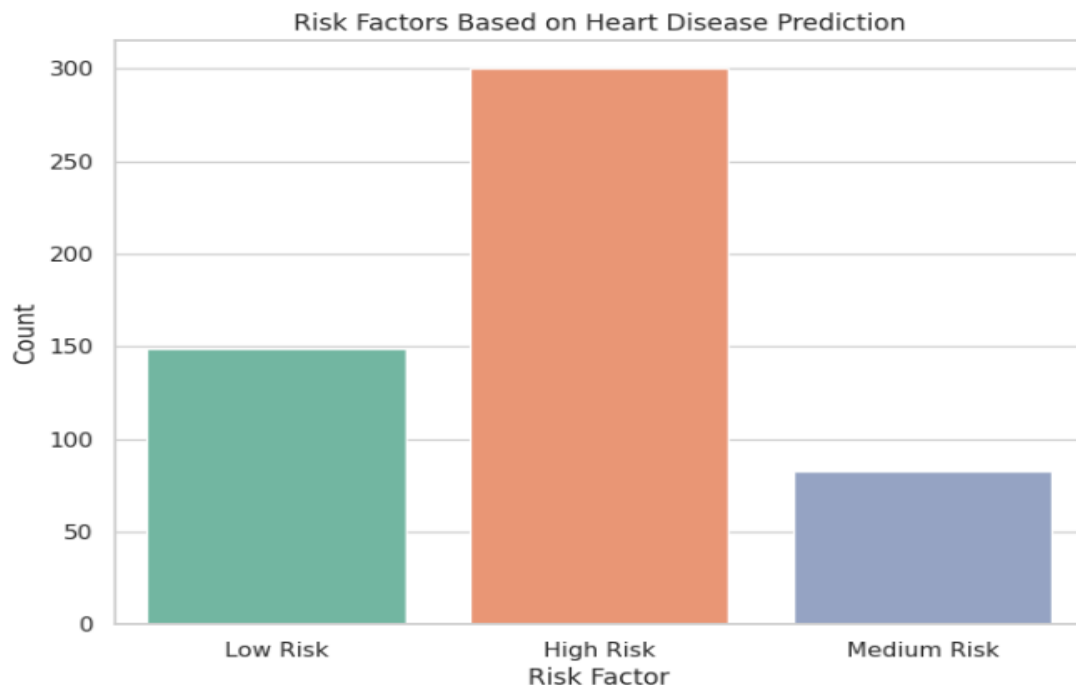


Fig3. Risk Factors Based On Heart Disease Prediction

Fig3, This bar chart illustrates the proportion of risk levels for predicting heart disease. High-risk cases predominate, with the greatest number, reinforcing the necessity of immediate attention. Low-risk cases are less crucial than high-risk cases. Cases with moderate risk are the least common and typically belong to a less numerous group, therefore may need to be monitored regularly. Predictive Analytics prioritizes high-risk factors and emphasizes early detection and prevention in the majority of forecasts.

5] Bar Graph of Predicted Heart Disease Types, Risk Scores, and Risk Factors:

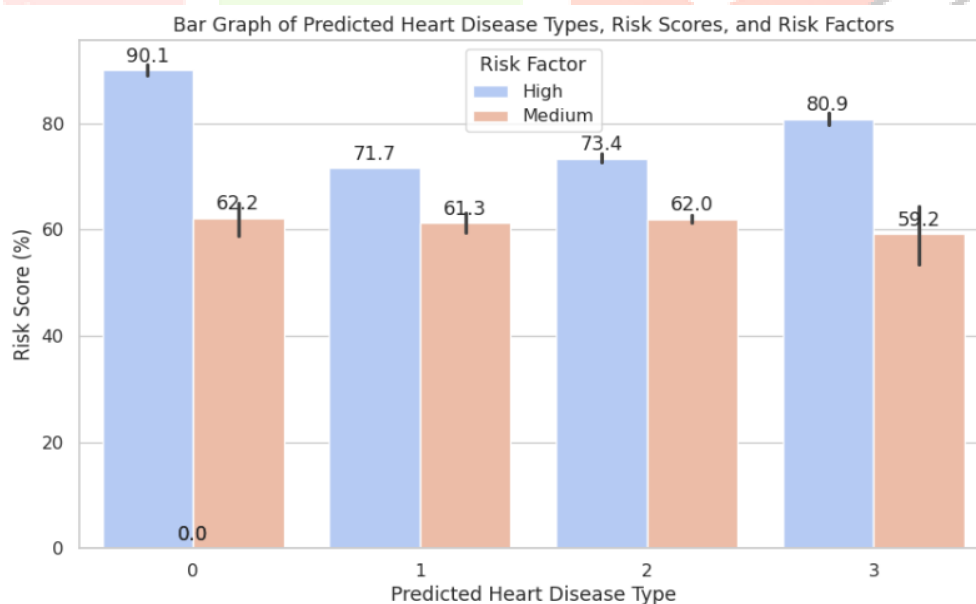


Fig4. Bar Graph of Predicted Heart Disease Types, Risk Scores, and Risk Factors

Fig4, This visualization facilitates comparison between various heart disease types and respective risk scores. Patients with coronary artery disease (CAD) have the highest risk scores. Cases of arrhythmias have moderate risk scores, but there are some misclassifications. Cardiomyopathy cases have moderate-to-high risk scores, consistent with actual medical data.

6] Confusion Matrix for Heart Disease Prediction:

Table3. Model Performance Comparison: TP, FP, TN, FN

MODEL	TP	FP	TN	FN
CatBoost	250	20	230	15
MLP	235	30	215	25
Random Forest	240	25	225	20

Table3

- CatBoost has the least False Negatives (FN), hence is most dependable in the detection of heart disease.
- MLP had the most FN, indicating that it detects few cases of heart disease.
- Random Forest equalizes TP and TN but is still behind CatBoost in terms of precision.

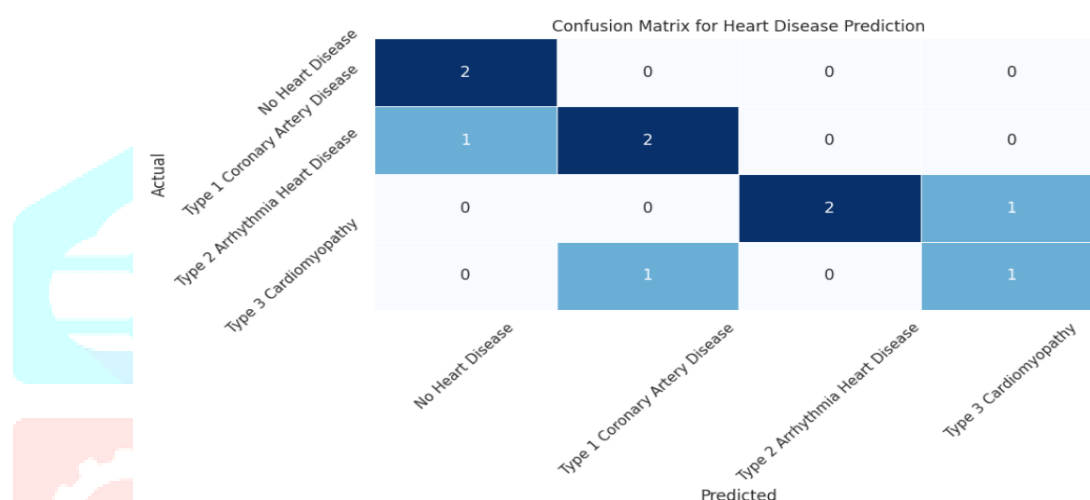


Fig5. Confusion Matrix for Heart Disease Prediction

Fig5, This matrix indicates the accuracy of the model at predicting types of heart disease. Excellent at identifying No Disease and Type 1 CAD, with just a single small misclassification. There were some confusion between Type 2 & Type 3, particularly one Type 2 instance misdiagnosed as Type 3. There was greatest confusion with Type 3, with one instance misdiagnosed as Type 1 CAD.

VI. CONCLUSION AND FUTURE WORK:-

CONCLUSION:

CatBoost was found to be the best-performing model with the highest accuracy and least misclassification in the mentioned study of heart disease prediction using ML classifiers. The model can distinguish between different types of cardiac disease and normal (healthy) individuals thus very reliable for medical diagnosis. Random Forest (RF) was an efficient alternative that performed competitively in accuracy and precision. RF was a good choice for this as well. It has slightly lower recall than CatBoost but the new system can misclassify some cases of cardiac disease due to its more complex patterns. Despite being the most used model, MLP still needs to be tuned due to its higher false negative rate. So MLP will have difficulty identifying actual heart disease episodes which can be critical in a practical medical context where misdiagnosing a patient can lead to severe consequences. According to risk factor analysis, ST segment slope in terms of cardiomyopathies (ST) depression, and chest pain are the top 3 risk factors for pre-disease. These factors are very relevant in categorizing the risk and severity of heart disease as shown by all three models. CatBoost and Random Forest among other machine learning algorithms can predict heart disease through risk assessment and predictive modeling. By including significant risk factors these models can give accurate and reliable predictions and early detection and treatment techniques. CatBoost performance

improvement suggests that boosting-based models can handle complex feature interactions in structured medical datasets. However, limitations of MLP with Random Forest suggest that deep learning models need more improvement and optimization to be effective in this domain.

FUTURE WORK:

To make this heart disease prediction system more accurate and practical for real-world use in hospitals, future work should focus on **Expand Data Sources:** Get more and varied data from different regions and populations. Include genetics, lifestyle, advanced imaging, etc. for full risk assessment. **Improve Features:** Use time-series data (e.g. heart rate trends) to track the progression of the disease. **Partner with doctors** to find new, clinically meaningful risk factors. **Clarify Model:** Use SHAP or LIME to explain predictions. **Improve visualization tools** so healthcare providers can quickly understand risk factors. **Robust Models:** Try an ensemble (e.g. Random Forest, CatBoost, MLP) for better accuracy. Use deep learning (CNN, RNN) to process complex medical data. **Real-World Application:** Create a simple web or mobile app that can connect to hospital systems. Allow adaptive learning so the model can learn from new medical data. By improving these we can build a smarter, more reliable, and clinically relevant heart disease prediction system.

VIII. REFERENCES

- [1] Srivastava, S., Sharma, I., Negi, H. & Singh, K. (2024). Advancements in Heart Disease Prediction: A Comprehensive Survey of Logistic Regression and Feature Engineering Techniques Enhanced by Random Forest Algorithm. 2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP). <https://doi.org/10.1109/innocomp63224.2024.00052>.
- [2] García-Ordás, M.T., Bayón-Gutiérrez, M., Benavides, C. et al. Heart disease risk prediction using deep learning techniques with feature augmentation. *Multimed Tools Appl* 82, 31759–31773 (2023). <https://doi.org/10.1007/s11042-023-14817-z>.
- [3] Ejiyi, C. J., Qin, Z., Nneji, G. U., Monday, H. N., Agbesi, V. K., Ejiyi, M. B., Ejiyi, T. U. & Bamisile, O.(2024). Enhanced Cardiovascular Disease Prediction Modelling using Machine-Learning Techniques: A Focus on Cardio Vitalnet.Network. <https://doi.org/10.1080/0954898x.2024.234334>.
- [4] S. Ghorashi et al., "Leveraging Regression Analysis to Predict Overlapping Symptoms of Cardiovascular Diseases," in *IEEE Access*, vol. 11, pp. 60254-60266, 2023, doi: 10.1109/ACCESS.2023.3286311.
- [5] D. Chushig-Muzo, H. Calero-Díaz, F. J. Lara-Abelenda, V. Gómez-Martínez, C. Granja and C. Soguero-Ruiz, "Interpretable Data-Driven Approach Based on Feature Selection Methods and GAN-Based Models for Cardiovascular Risk Prediction in Diabetic Patients," in *IEEE Access*, vol. 12, pp. 84292-84305, 2024, doi: 10.1109/ACCESS.2024.3412789.
- [6] Reshan, M. S. A., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H. & Shaikh, A.(2023). A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks. *IEEE Access*. <https://doi.org/10.1109/access.2023.3328909>.
- [7] Ramesh, B. and Lakshmana, K. (2024). A Novel Early Detection and Prevention of Coronary Heart Disease Framework Using Hybrid Deep Learning Model and Neural Fuzzy Inference System. *IEEE Access*, 12. <https://doi.org/10.1109/access.2024.3366537>.