Cardiovascular Disease Prediction with Machine Learning-Based Clinical Decision Support System

Naga Satwika Kakarla
Computer Sci.
Kent State University
nkakarl1@kent.edu

Tejasree Kilari
Computer Sci.
Kent state University
tkilari@kent.edu

Lakshmi Durga Bhavani Meesala
Computer Sci.
Kent state University
lmeesal1@kent.edu

Kamaluddin Syed
Computer Sci.
Kent state University
ksyed3@kent.edu

Shamshiksai Peddoju
Computer Sci.
Kent state University
speddoju@kent.edu

***Abstract*—Cardiovascular disease, particularly heart disease, is a leading cause of morbidity and mortality worldwide. Early detection and accurate prediction of heart disease risk factors are central to the implementation of preventive measures and individual health care. This study investigates the application of machine learning algorithms to the prediction of heart disease based on a comprehensive analysis of various clinical data. The data used in this study include a wide range of demographic, clinical, and lifestyle factors collected from a large patient population. A variety of machine learning algorithms are used to develop predictive models, including but not limited to decision trees, support vector machines, random forests, and neural networks. Feature selection techniques and hyperparameter tuning are used to improve model performance and interpretability. The study evaluates the predictive accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC-ROC) of each model to determine the most effective algorithm for predicting heart disease. In addition, interpretable analyzes are performed to increase the transparency of the models and integrate them into clinical practice. The results of this study will contribute to ongoing efforts to improve methods of heart disease risk assessment. Identifying robust machine learning models to predict heart disease will allow significant improvements in early diagnosis and optimization of prevention strategies, thus reducing the public health burden of cardiovascular disease. This study lays the groundwork for future work to integrate machine learning algorithms into clinical decision support systems to achieve more efficient and personalized patient care**

I. Introduction

Imagine a super-smart tool that helps doctors predict heart issues early. This tool, using fancy computer learning, looks at lots of info about patients and spots signs of possible heart troubles. It's like a health guardian angel! By teaming up computer smarts with doctor know-how, it makes heart care better. Before someone even feels sick, this system can say, "Hey, watch out for this!" This way, doctors can act quickly and give personalized advice. It's like having a heads-up on your heart health, making sure problems are caught early, and everyone stays healthier. This cool tech isn't just about fixing hearts; it's about keeping them happy and strong from the start

II. Description

Heart disease is a global health concern, with a rising annual impact due to multiple contributing risk factors, including diabetes, high blood pressure, and elevated cholesterol levels. Accurate prediction of heart disease is vital for averting life-threatening situations. To achieve more precise identification, we are employing a range of machine learning algorithms, such as Logistic Regression, Random Forest, SVM, AdaBoost, K-Nearest Neighbor Classifiers, Naive Bayes, and decision trees. These algorithms are pivotal in enhancing our predictive model's effectiveness, with the ultimate goal of improving heart disease identification and patient outcomes.

III. Literature Review

The authors in paper [1] explains the use of IoT in healthcare for remote patient monitoring and real-time diagnosis. They also explains the importance of precision agriculture.

The paper [2] explains using ML on blood test data to predict COVID-19 mortality. The demerits found in paper are Reduced F1 score over time, inconsistency, overfitting, difficulty in model interpretability.

The paper [3] explains the use of machine learning for accurate coronary artery disease prediction. The methods proposed by authors are best for handling imbalanced data but disadvantages found are Over fitting, noise due to random oversampling, runtime complexity from SMOTE.

The authors in paper [4] proposed methods to identifying multiple chronic diseases efficiently using Machine learning techniques. It provides best early disease prediction solution but have disadvantage of getting inaccurate results from user.

The paper [5] explains the method for Early heart abnormality diagnosis using patient clinical features. But authors used very small dataset which may lead to inaccuracy in results.

The paper [6] explains how Machine learning can be used for early and non-invasive heart disease diagnosis. But methods proposed by authors are bad at handling complex data and has potential model incompatibility

IV. Study Approach

The Goals described by studying various papers : Engaging in qualitative research, our endeavor involved a comprehensive exploration of relevant literature, providing us with a nuanced understanding of our objectives. Our articulated goals underscore our commitment to advancing the field: Implementation of Advanced Machine Learning Algorithms:Our primary focus was on elevating the precision of heart disease prediction by incorporating state-of-the-art Machine Learning Algorithms, surpassing the efficacy of prior projects. Enhancement of Dataset Quality: Recognizing the limitations of smaller datasets in previous projects, we diligently curated a more extensive and up-to-date dataset, enriching it with additional attribute information to bolster the robustness of our analyses. Comprehensive Data Visualization: Departing from conventional approaches, we employed data visualization techniques to elucidate the intricate relationships between features and the target variable. Unlike other studies that selectively displayed feature-target relations, we comprehensively presented all features, fostering a holistic understanding. Identification of Key Features: Leveraging diverse methodologies, we discerned the crucial features influencing the target variable, employing a discerning approach to neglect less impactful attributes. This strategic feature selection enhances the interpretability and efficacy of our predictive model, contributing to the refinement of heart disease prognostication.

V. Methodology

The methodologies used in cardiovascular disease prediction using machine learning classifiers involve the integration of various technologies and systems. Here we used steps data acquisition, data cleaning, data visualization, data preprocessing on the selected dataset. Then selection of machine learning algorithms is based on the prepared data, adapting to the characteristics of cardiovascular health prediction.

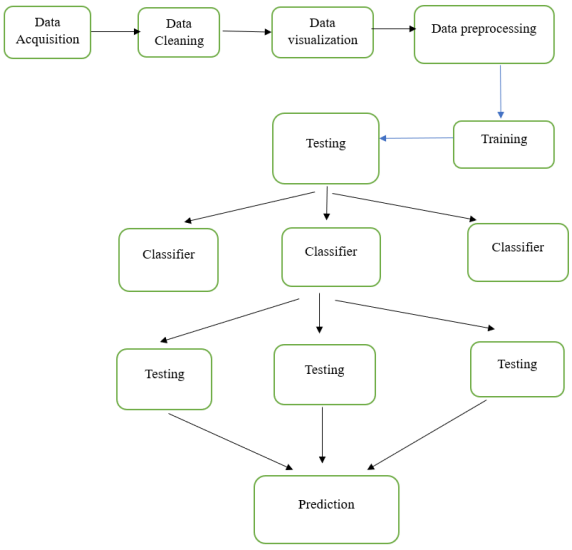

Fig. 1. Methodology

A. Data Acquisition

The dataset used for prediction of heart disease is sourced from the UCI repository and is accessible at https://archive.ics.uci.edu/dataset/45/heart+disease. Comprising 918 observations, it encompasses 12 attributes, with one designated as the target field, "heart disease," indicating the presence of the condition. This integer-valued target field distinguishes between no heart disease (labeled as 0) and the presence of heart disease (labeled as 1). The dataset contains 11 predictive features that aid in determining the likelihood of a patient having heart disease.

```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

Fig. 2. Attributes of dataset

B. Data Cleaning

We have thoroughly checked the data set for empty values and duplicate values also we clean the data and make it more flexible for the further process. We have used **df.isnull().sum()** to return the sum of missing values in the dataset. From this we found that there are no null values in our dataset.

C. Data Visualization

Visualization is a powerful tool for simplifying complex data and uncovering patterns that aid in informed decision-making for treatment, prevention, and research related to the target variable. By employing visualization, we can determine the most influential features on the target variable and identify any outliers that may need to be removed to improve accuracy. In our project, we used pairplots to visualize the relationships between the features and the target variable. These visualizations helped us identify an outlier in the form of a resting blood pressure value of zero, prompting us to remove it from the dataset for improved data quality.
We also showed the relationship between other features and target by plotting the visualization for each feature and target.

```
[ ]  df.isnull().sum()
```

```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

Fig. 3. Null values in dataset

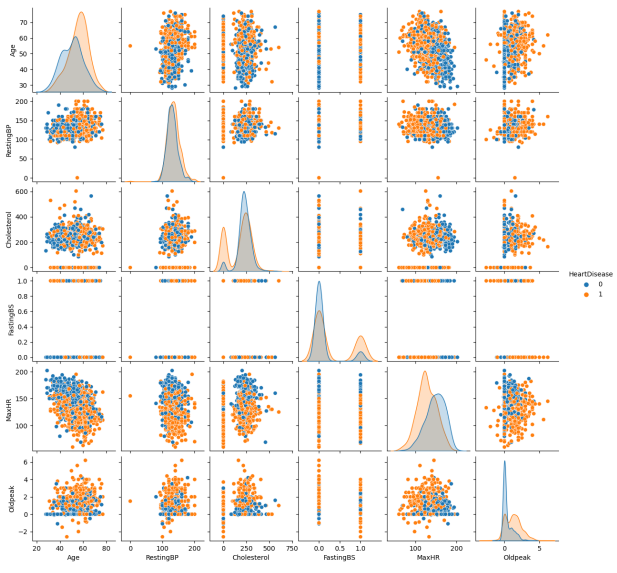

Fig. 4. Visulaization of target-feature relationship

Fig. 5. Relationship between heart disease and Sex(Male or Female)

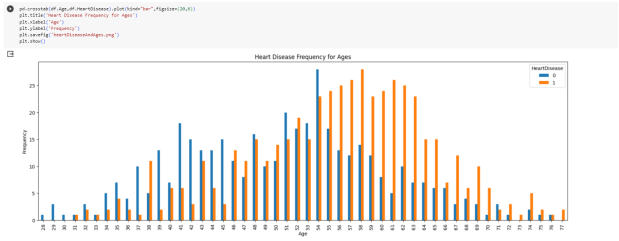

Fig. 6. Relationship between target (heart disease) and Age

D. Data Preprocessing

Data preprocessing is a crucial initial step in data analysis and machine learning. Its primary purpose is to enhance the quality and usability of raw data through processes like data cleaning, transformation, and structuring. This involves tasks such as handling missing data, dealing with outliers, converting, and encoding data, creating new features, and dividing the data into training and testing sets. In more complex situations, you might also need to reduce the dimensionality of the data or integrate data from multiple sources. Visualization is often used to gain insights into the data, and documenting the preprocessing steps is vital for transparency and reproducibility. The data preprocessing process is not a one-time task but rather an iterative one that varies depending on the specific data and project objectives. It plays a fundamental role in ensuring that the data is in the right format and quality for accurate analysis and successful machine learning. In this context, data preprocessing often involves converting categorical features into numerical ones to make them compatible with machine learning models. For example, when dealing with a feature like ”exercise angina” that has values ’Y’ for yes and ’N’ for no, it’s common to replace ’Y’ with 1 and ’N’ with 0. This conversion transforms the categorical feature into a binary one, where 1 signifies the presence of exercise-induced angina (True) and 0 indicates its absence (False).

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	S
0	40	M	ATA	140	289	0	Normal	172	N	0.000000	
1	49	F	NAP	160	180	0	Normal	156	N	1.000000	
2	37	M	ATA	130	283	0	ST	98	N	0.000000	
3	48	F	ASY	138	214	0	Normal	108	Y	1.500000	
4	54	M	NAP	150	195	0	Normal	122	N	0.000000	
5	39	M	NAP	120	339	0	Normal	170	N	0.000000	
6	45	F	ATA	130	237	0	Normal	170	N	0.000000	
7	54	M	ATA	110	208	0	Normal	142	N	0.000000	
8	37	M	ASY	140	207	0	Normal	130	Y	1.500000	
9	48	F	ATA	120	284	0	Normal	120	N	0.000000	

Fig. 7. Showing features with heatmaps

Here in preprocessing we use Label Encoder to convert Categorical values to numerical values.

E. Data Manipulation and Exploration

Data manipulation and exploration are essential data analysis processes. Data manipulation involves tasks such as cleaning, transforming, and restructuring data to prepare it for

```
[ ] def label_encode_cat_features(data, cat_features):
        '''
        Given a dataframe and its categorical features, this function returns label-encoded dataframe
        '''

        label_encoder = LabelEncoder()
        data_encoded = data.copy()

        for col in cat_features:
            data_encoded[col] = label_encoder.fit_transform(data[col])

        data = data_encoded

        return data
```

Fig. 8. Function for label encoding

analysis. This includes handling missing values, encoding categorical variables, and feature engineering. Data exploration, on the other hand, involves visualizing and summarizing data to uncover patterns, relationships, and insights.

Techniques like histograms, scatter plots, and statistical measures help in this process. These activities aid in understanding the data's characteristics, identifying outliers, and preparing it for modeling or analysis, ultimately leading to more informed decision-making and valuable insights from the data.

Fig. 9. Histogram of features

F. Machine learning Classifiers

We used Scikit-learn Classifiers and boosted tree classifiers to train the model.

VI. Analysis and Results

To evaluate the performance of the machine learning classifiers, we used the following metrics:

1) Accuracy: The proportion of correctly predicted cases.
2) ROC_AUC curve: The area under the receiver operating characteristic curve, which measures the classifier's ability to distinguish between positive and negative cases.
3) Recall: The proportion of actual positive cases that are correctly predicted.
4) Precision: The proportion of predicted positive cases that are actually positive.
5) F1 score: The harmonic mean of precision and recall.

The results are:

The generated confusion matrix plots for classifiers are shown in Fig.11.

The Naive Bayes and Quadratic Discriminant Analysis Classifier achieved the best performance on the test set, with accuracy score of 86.96 percent.These classifiers also achieved

	Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
8	Naive Bayes	86.960000	0.940000	0.900000	0.870000	0.890000
10	Quadratic DA	86.960000	0.950000	0.920000	0.860000	0.890000
7	Gradient Boosting	86.520000	0.940000	0.920000	0.850000	0.890000
11	Neural Net	85.220000	0.920000	0.870000	0.870000	0.870000
5	Random Forest	84.780000	0.920000	0.900000	0.840000	0.870000
9	Linear DA	84.780000	0.920000	0.880000	0.860000	0.870000
6	AdaBoost	84.350000	0.920000	0.870000	0.860000	0.860000
3	Nu SVC	83.910000	0.910000	0.910000	0.830000	0.870000
0	Logistic Regression	83.480000	0.920000	0.890000	0.830000	0.860000
4	Decision Tree	78.260000	0.770000	0.84000	0.790000	0.810000
2	Support Vectors	73.040000	0.800000	0.780000	0.7600	0.770000
1	Nearest Neighbors	65.650000	0.6	0	0.77000	0.66

Fig. 10. Performance metrics when data is split into 75% for training and 25% for testing

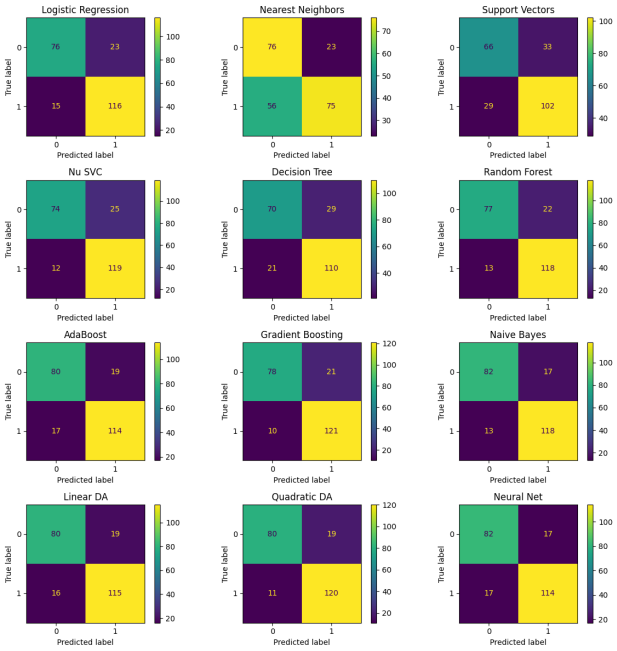

Fig. 11. Performance metrics

high ROC_AUC scores, indicating that they are good at distinguishing between positive and negative case.

The other classifiers also performed well, with accuracy scores in range 80-60 percent.

We also evaluated the performance of three modern boosted tree algorithms: CatBoost, XGBoost, and LightGBM. These algorithms are known to be very effective for a variety of machine learning tasks, including classification.

	Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
0	Catboost	87.830000	0.940000	0.900000	0.890000	0.890000
1	xgbbost	83.480000	0.920000	0.890000	0.830000	0.860000
2	light GBM	83.040000	0.930000	0.890000	0.820000	0.860000

Fig. 12. Performance metrics for modern classifiers

The following figure shows the confusion matrices for the boosted tree algorithms on the test set:

The CatBoost Classifier with accuracy score of 87.83 %. Other classifiers also performed well with accuracy scores above 80%.

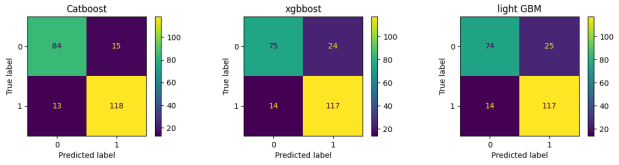

Fig. 13. Confusion matrix

A. Comparison of scikit-learn classifiers with boost tree algorithms

The boosted tree algorithms outperformed the scikit-learn classifiers on the dataset. The CatBoost Classifier achieved accuracy score of 87%, which is higher than the highest accuracy score achieved by the scikit-learn classifiers.

VII. Analysis of results

We initially divided the dataset into 75% and 25% for training and testing our machine learning model. Here Quadratic Discriminant Analysis (Quadratic DA) classifier performed best in terms of Accuracy and Recall. Since we are considering Recall as our main criteria for selecting the best classifier, We can say that Quadratic DA classifier performed the best.

Then we split the dataset into 60% and 40% for training and testing to check how our classifiers work on different splits of data. Here we can say that Quadratic DA classifier still proved to be best in terms of accuracy and Recall.

Then we used Advanced Machine learning classifiers . We can see that catboost classifier performed the best in terms of accuracy, precision, Recall and F1 score. Since we are considering Recall as our main criteria in selecting the best performing classifier, we can say that Catboost classifier is best performing classifier.

VIII. Summary

Predicting cardiovascular disease (CVD) using machine learning algorithms entails harnessing data-driven models to scrutinize patient data and assess CVD risk. Our primary objectives are to leverage comprehensive data visualization to unveil feature-target relationships and identify crucial features, enhance dataset quality by incorporating more attribute information and utilizing the latest available dataset, and deploy advanced machine learning algorithms to refine heart disease prediction accuracy.

We employed the most recent and updated dataset with a broader range of attributes from the UCI website to achieve superior results. Additionally, we utilized both traditional and advanced machine learning classifiers to elevate overall model performance.

By comparing both approaches, we achieved enhanced and rapid accuracy with Boosted Tree Classifiers (e.g., Catboost, Xgboost, Light GBM).

Catboost emerged as the superior classifier, demonstrating the highest accuracy and outperforming in terms of recall and precision. Furthermore, Catboost is 30 to 60 times faster than XGBoost and LightGBM.

CatBoost facilitates in-depth analysis by enabling a straightforward focus on a single feature's contribution to the model, rather than a specific value of a specific feature. Catboost consistently outperformed all other machine learning classifiers, achieving an overall accuracy of approximately 87

CatBoost is a robust machine learning algorithm that excels in various domains, including ease of use, categorical feature handling, scalability, accuracy, overfitting prevention, prediction speed, and feature importance analysis.

These factors make Catboost the most suitable choice for our dataset. Catboost maintains its efficiency even when data size increases

IX. Acknowledgements

Dr.Samba helped to learn the process of the capstone project which advanced aspects.We are grateful to him and it is an nice learning experience. We are writing to express our sincere gratitude for your invaluable guidance and support throughout our final project. Your expertise, dedication, and unwavering encouragement have played a crucial role in the successful completion of our project. From the early stages of project conception to the final stages of presentation, you have consistently provided us with insightful feedback, constructive criticism, and unwavering encouragement. Your ability to identify areas for improvement and guide us towards solutions has been instrumental in refining our project's scope and methodology. We are particularly grateful for your patience and willingness to address our questions and concerns, no matter how small or insignificant they may have seemed. Your availability and willingness to go the extra mile have been invaluable in helping to overcome challenges and maintain a positive and productive mindset throughout the project.

References

[1] C. Chakraborty and A. Kishor, "Real-Time Cloud-Based Patient-Centric Monitoring Using Computational Health Systems," in IEEE Transactions on Computational Social Systems, vol. 9, no. 6, pp. 1613-1623, Dec. 2022, doi: 10.1109/TCSS.2022.3170375.

[2] Karthikeyan Akshaya, Garg Akshit, Vinod P. K., Priyakumar U. Deva,"Machine Learning Based Clinical Decision Support System for Early COVID-19 Mortality Prediction " in Frontiers in Public Health , vol. 9 ,2021,https://www.frontiersin.org/articles/10.3389/fpubh.2021.626697,doi: 10.3389/fpubh.2021.626697.

[3] K. N. Devi, S. Suruthi and S. Shanthi, "Coronary Artery Disease prediction using Machine Learning Techniques," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 1029-1034, doi: 10.1109/ICACCS54159.2022.9785140.

[4] M. Manwal, D. Aswal and V. Tewari, "Machine Learning Based StreamLit API Multi-Disease Detection," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 266-272, doi: 10.1109/ICIRCA57980.2023.10220660.

[5] M. R. Islam, M. Durul Hoda, M. A. Rashid, S. Alam Suha and M. T. Islam Miya, "Data-Driven Heart Disease Prediction by Ensemble Feature Selection and Machine Learning Techniques," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 575-580, doi: 10.1109/ICCIT57492.2022.10054998.

[6] R. Mishra and S. K. Gupta, "Optimization Accuracy of Heart Disease Diagnosis using Machine Learning Approach," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/GCAT55367.2022.9971872.

[7] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.