CardioCare: Cardiovascular Anomaly Detection

Arpit Patil
Department of Computer Science (ML)
Stevens Institue of Technology
Hoboken, NJ, United States
apatil45@stevens.edu

Yash Tanaji Bhosale
Department of Mathematical Science
Stevens Institue of Technology
Hoboken, NJ, United States
ybhosale@stevens.edu

Neel Patel
Department of Computer Science
Stevens Institue of Technology
Hoboken, NJ, United States
npatel21@stevens.edu

***Abstract*—This project focuses on diagnosis of cardiovascular anomalies which is important for giving an early treatment response using comprehensive set of Machine Learning Algorithms. The project will be focused on classifying cardiovascular anomalies based on the following machine learning algorithms: CART, Naïve Bayes, Random Forest, and Convolutional Neural Networks. The data preprocessing will involve normalization, feature selection, and dimensionality reduction to assure better performance. The initial results using CART and NB provided the baseline in terms of accuracy and efficiency, while Random Forest and CNN show promise in capturing the complex patterns in this dataset. These algorithms will be compared in the final analysis to find which model best suits anomaly detection and thus helps address the real-world challenges in cardiovascular diagnostics.**

I. Introduction

Cardiovascular Diseases (CVDs) are the leading cause of death, thus continuing to have high mortality rates in the world. The key for better outcomes and a reduction in the burden of these diseases is early detection and precise diagnosis of cardiovascular anomalies. Traditional diagnostic methodologies include ECG, imaging, and stress tests. These detection tools take time and need expert professionals. While the volume of patient data keeps increasing, it is now an opportunity for machine learning to automate such detection mechanisms and make the processes effective in arriving at quicker or probably even more accurate results. The possibility of early detection of anomalies such as arrhythmias, ischemic condi- tions, and dysfunction of heart valves will greatly enhance clinical decision-making and patient care. The aim of this project is to introduce a new approach in the detection of cardiovascular anomalies by using different machine learning algorithms, namely, CART, NB, Random Forest, and CNN. Each of these algorithms has certain ad- vantages, and this study tries to find their efficiencies for the detection of cardiovascular anomalies based on clinical data. The main aim is to create a strong, efficient, and scalable model for categorizing normal versus abnormal heart conditions that can assist in early diagnosis.

The project systematically covers the steps, starting with data preprocessing: cleaning, handling missing values, and normalization of the dataset. Moreover, feature extraction techniques such as PCA and statistical methods are used to enhance model performance by focusing on the most relevant features for anomaly detection. Real-world cardiovascular datasets, comprising patient health metrics along with their diagnostic records, are used for training and testing.

The first baseline model utilized the CART algorithm, which was giving an accuracy of about 80 around the midpoint of the project. The first model in most cases normally tends to provide a benchmark with which more advanced algorithms can be compared. Naïve Bayes is a probabilistic model and hence expected to perform well with smaller and less complex data, making it suitable for the classification tasks at the initial stage. Random Forest is also an ensemble method based on decision trees that is robust and copes well with big, noisy datasets. Finally, deep learning methods are applied, especially Convolutional Neural Networks, to model more elaborate nonlinear relationships within the data when time-series signals are analyzed, such as ECGs or any other form of medical imaging data.

This we will further in the final stages of the project, where implementation and performance evaluation of all the algorithms will be performed on the same dataset: namely, CART, Naïve Bayes, Random Forest, and CNN. We will compare the performance based on the classification accuracy, precision, recall, and F1-score for each model, along with computational efficiency. This will, in turn, enable us to determine which one performs best for cardiovascular anomaly detection and give us an insight into the suitability of the model for clinical use. The study will prove how machine learning algorithms can change the face of cardiovascular diagnosis. These methods will contribute to automatically helping medical experts detect anomalies quicker and make more accurate diagnoses for better patient outcomes, thus reducing the burden of cardio- vascular diseases worldwide.

II. Related Work

In recent years, cardiovascular anomaly detection using machine learning has emerged as a pivotal area of research, addressing the challenges of early diagnosis and precise monitoring of heart diseases. This section reviews existing work related to anomaly detection in cardiovascular studies and highlights the role of the algorithms utilized in this project.

A. *Tradition Machine Learning Algorithms*

Traditional machine learning had formed the backbone in cardiovascular anomaly detection, owing to its scalable and interpretable models. Decision trees form the basis for Classification and Regression Trees; these methods optimally split datasets into homogenous groups based on cardiovascular risk factors. However, their use of recursive partitioning may lead them to overfitting. Naive Bayes represents a probabilistic classifier based on Bayes' theorem. It works, especially in cases of low dimensional feature space due to feature independence. Random Forest ameliorates basic pitfalls of CART since it is an ensemble of decision trees. Predictions of a great number of trees are averaged in order to enhance generalization and robustness. Although traditionally having application-based relevancy with image data, recent adaptations of CNNs that analyze ECG signals have taken advantage of the ability to extract spatial hierarchies and patterns. While these traditional methods have shown reliable performance, in complex, high-dimensional datasets of cardiovascular data, they may require very extensive manual feature engineering.

B. *Hybrid and Ensemble Methods*

Hybrid and ensemble methods combine the best of various algorithms to come up with an improved model performance for diagnosing cardiovascular anomalies. Strate- gies such as stacking and bagging leverage and build on diverse algorithms in order to create a single, unified output. Thus, for example, the Random Forest ensemble by bagging sums up the output of several CART models to come up with a more robust classifier. Similarly, hybrid models that incorporate probabilistic algorithms, like NB, into deep learning frameworks such as CNNs exploit the advantages of both simplicity in preliminary classification and sophisticated pattern recognition. These techniques generally perform well in handling heterogeneous datasets and imbalanced classes- often found in medical applications-thus performing better with increased robustness to noise compared to approaches that use a single model.

C. *Transfer Learning with Transformer Models*

Recently, transformer models initially targeted for natural language processing have become very popular for transfer learning in cardiovascular anomaly detection. In particular, pretraining transformer models like BERT and Vision Transformers allow one to extract high-order features of cardiovascular data represented by ECG signals and medical images. Fine-tuning these models on domain-specific datasets can allow researchers to leverage learned representations from vast unrelated datasets, strongly saving training time and boosting the accuracy of the models. Transformer-based architectures achieve state-of-the-art performance in capturing temporal dependencies and, therefore, are especially suitable for sequential data, such as ECG signals. In the domain of cardiovascular anomaly detection, their application is at an early stage but forms a promising direction of integration of state-of-the-art AI techniques into clinical diagnostics.

III. OUR SOLUTION

This section elaborates on our solution to the problem of Cardiovascular Anomaly Detection by detailing the dataset, chosen machine learning algorithms, and implementation approach.

A. *Description of Dataset*

There are two sections to the PTBDB datasets: abnormal and normal. Each dataset has one target column that indicates the categorization and 187 numerical features that represent samples. 10,506 records in the aberrant dataset have the target labeled as 1.0, whereas 4,046 entries in the normal dataset have the target labeled as 0.0. Both datasets offer normalized numerical values and have the same structure.

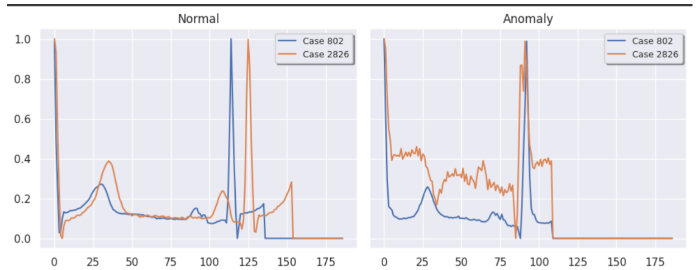


Fig. 1: Categorized Dataset

When the objective is to differentiate between normal and aberrant samples, these datasets are perfect for binary classification tasks. They require less preprocessing for model training and analysis because of their labeled targets and consistent format.

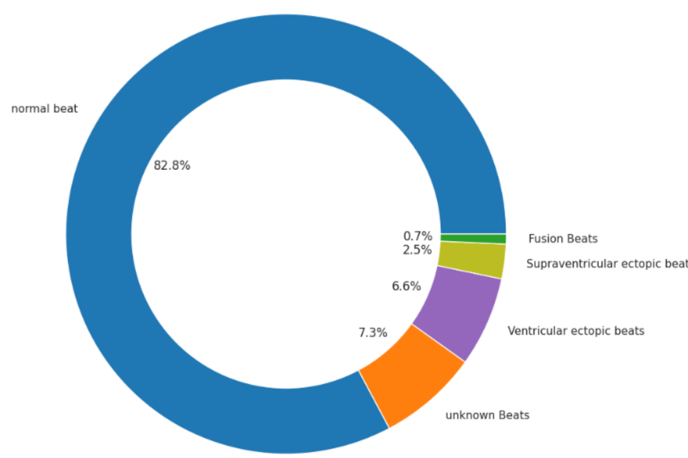


Fig. 2: Represntation of Classes

B. Exploratory Data Analysis (EDA)

In this project, a detailed Exploratory Data Analysis (EDA) was performed to understand the dataset and uncover important patterns. First, we generated an interactive correlation matrix for both the normal and anomaly datasets. This allowed us to visualize relationships between the features and identify which features are strongly correlated. Understanding these correlations helped us determine which features might have higher importance in distinguishing between normal and anomalous data.

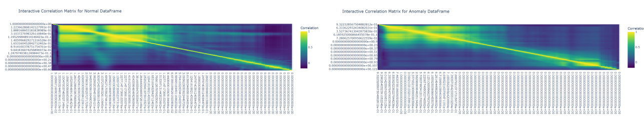


Fig. 3: Normal and Anomaly Correlation

Additionally, we created boxplots for selected features in the datasets. These visualizations highlighted the presence of outliers and provided insights into the spread and distribution of values. By comparing the feature distributions for normal and anomaly data, we were able to detect significant variations, which played a critical role in preprocessing and feature selection. The EDA provided valuable insights that shaped our data processing and modeling strategy.

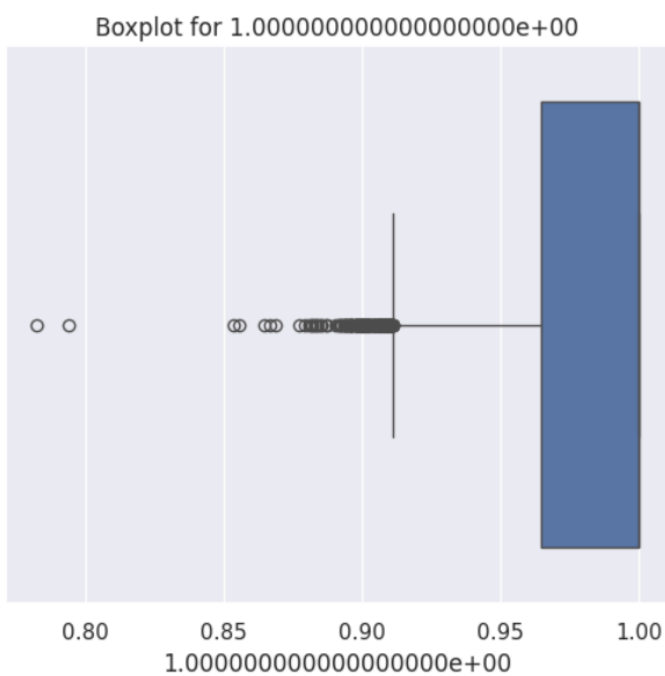


Fig. 4: Sample Boxplot

PCA was applied as a feature reduction technique to reduce the complexity of the dataset, retaining as much variance as possible. The dataset initially comprised high-dimensional features, which normally results in computational inefficiency and may lead to overfitting. PCA transforms data into a lower dimensional space by extracting the principal components that capture the most significant amount of variance in the data. It retained 95 of the total variance, hence preserving important information while reducing redundancy. In this way, it allowed a better visualization of the distribution and improved the efficiency of the models used afterwards.

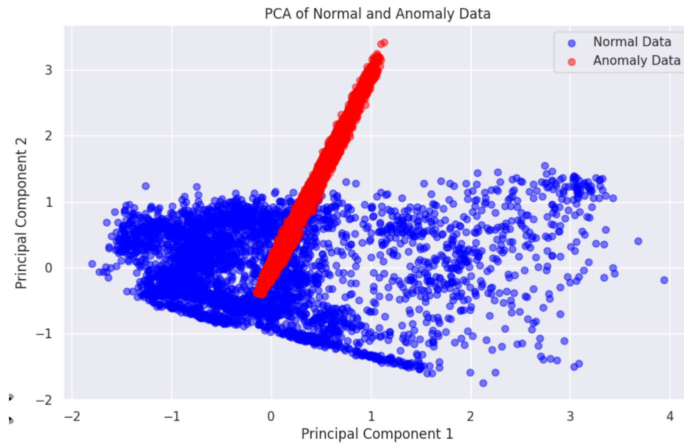
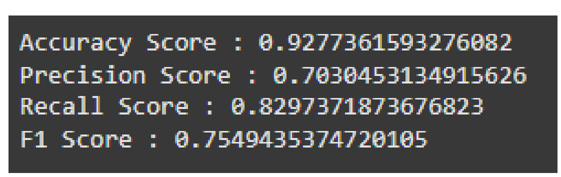


Fig. 5: Principal Component Analysis

C. Machine Learning Algorithms

Our approach evaluates multiple machine learning algorithms to identify the most effective model for Cardiovascular Anomaly Detection. We selected four models, each chosen for specific strengths that address the unique challenges of analyzing normal and abnormal anomalies.

- **CART (Classification and Regression Trees):** CART (Classification and Regression Trees): A basic component of classic machine learning, CART is an intuitive and simple method for decision-making. In this work, the CART algorithm demonstrated a very good accuracy of 92.77, with a precision of 70.30, recall of 82.97, and an F1 score of 75.49, proving to be efficient in the detection of cardiovascular anomalies. The algorithm iteratively divides the dataset into smaller, more homogeneous groups based on features like blood pressure or heart rate, which results in a tree structure where leaves represent final classifications and branches denote decision rules. Its interpretability and simplicity make it a very useful tool in medical applications. However, the relatively lower precision score indicates a tendency for false positives, suggesting that some predictions may require additional verification. Pruning strategies may be necessary to avoid overfitting, especially in high-dimensional datasets. Despite these challenges, CART remains a very reliable algorithm and often forms the foundation for ensemble methods such as Random Forest to enhance accuracy and robustness.



```
Accuracy Score : 0.9277361593276082
Precision Score : 0.7030453134915626
Recall Score : 0.8297371873676823
F1 Score : 0.7549435374720105
```

Fig. 6: Cart Metrics

- **Naive Bayes :** Naive Bayes is a simple probabilistic classifier based on Bayes' theorem with very low computational load. However, during the project, NB performed pretty poorly, reaching only 18.88 of accuracy, with precision at 33.57, recall at 45.87, and an F1 score at 17.79. The confusion matrix further shows how it fails to correctly classify instances among multiple classes, with significant misclassifications, especially in distinguishing between normal and anomalous data. NB assumes conditional independence among the features such as heart rate, cholesterol levels, and metrics of ECG signals, which may not be the case for complex cardiovascular datasets, where often there is a strong correlation among the features. While the simplicity and speed make it suitable for preliminary analysis or small datasets, the algorithm relies on the independence assumption, which limits its performance in medical applications. It could be combined with dimensionality reduction or feature selection to enhance its effectiveness in handling correlated features and improving predictive accuracy.

```
Accuracy Score : 0.18888178329983554
Precision Score : 0.33573706699674966
Recall Score : 0.45874728202303494
F1 Score : 0.177890660826155
```

Fig. 7: Naive Bayes Metrics

- **Convolutional Neural Networks (CNN) :** While CNNs were originally designed for image processing, they have been adapted with great success for the analysis of time-series data, including ECGs. In this project, the CNN showed a performance of 46 in general. It had high recall for anomalies at 87, but the precision for anomalies was comparatively low at 32, which means that there were more false positives. On normal data, the best performance of the CNN for precision was 86, but its recall was comparably low at 30. This suggests that the model has a bias toward anomalous class detection at the cost of normal class performance. CNNs are remarkable in feature learning from raw ECG signals by detecting important spatial patterns and hierarchical representations through convolutional layers, while dimensionality reduction helps improve computational efficiency through pooling layers. Limiting factors of the performance would include data imbalance challenges, or poor generalization after feature extraction. Pre-data augmentation or weighted loss can be considered to develop much better performance of a CNN in medical anomaly detection applications.

	precision	recall	f1-score	support
Normal	0.86	0.30	0.45	10505
Anomaly	0.32	0.87	0.47	4045
accuracy			0.46	14550
macro avg	0.59	0.59	0.46	14550
weighted avg	0.71	0.46	0.45	14550

Fig. 8: CNN Metrics

- **Random Forest :** The Random Forest ensemble learning technique builds a lot of decision trees during training and then combines the results in order to classify the inputs. In this project, Random Forest performed well: it achieved an accuracy of 97.69, precision of 91.16, recall of 87.17, and an F1 score of 89.07. The confusion matrix illustrates a well-performing Random Forest for multi-class classification tasks, with very few misclassifications for most of the classes, thus proving very robust in anomaly detection within the cardiovascular domain. Random Forest works exceptionally well with noisy or unbalanced datasets, making them suitable for medical applications where often redundant or irrelevant features are present that can distort your findings. By averaging predictions over its ensemble, Random Forest reduces the risk of overfitting-a common limitation of individual decision trees-thus improving generalization. The algorithm also offers valuable insights into feature importance that help clinicians identify predictive factors such as cholesterol levels or ECG wave amplitudes. Strong performance combined with interpretability makes Random Forest one of the most reliable choices for anomaly detection in healthcare.

```
Accuracy Score : 0.9769322126804312
Precision Score : 0.9116034120558927
Recall Score : 0.8717621064162857
F1 Score : 0.8907001620310337
```

Fig. 9: Random Forest Metrics

D. Implementation Details

- **Data Preprocessing:** Data Preprocessing: The PTBDB dataset was loaded and a good cleaning was done. All the missing values were taken care of here, hence ensuring the data quality. Features were brought into a comparable scale using the Min-Max scaling technique for models such as CART, Naive Bayes, and Random Forest. For the CNN model, z-score standardization was utilized to improve convergence during the training process. Feature correlation analysis was done to find out the relationship between features. Variance thresholds were used to remove redundant or irrelevant features. PCA was used to reduce the dimensions, choosing 95 of the variance to decrease complexity without losing critical information. The data was divided into 70 for training and 30 for testing, and all the necessary steps were taken to maintain class balance to ensure fair evaluation across all models.

```python
# Handle missing values (NaN)
# Check for missing values in both DataFrames
print("Missing values in Normal DataFrame:")
print(normal_df.isnull().sum())

print("\nMissing values in Anomaly DataFrame:")
print(anomaly_df.isnull().sum())

# Impute missing values with the mean of each column
imputer = SimpleImputer(strategy='mean')
normal_df_imputed = pd.DataFrame(imputer.fit_transform(normal_df), columns=normal_df.columns)
anomaly_df_imputed = pd.DataFrame(imputer.fit_transform(anomaly_df), columns=anomaly_df.columns)

# Confirm there are no missing values left
print("\nMissing values after imputation in Normal DataFrame:")
print(normal_df_imputed.isnull().sum())

print("\nMissing values after imputation in Anomaly DataFrame:")
print(anomaly_df_imputed.isnull().sum())
```

Fig. 10: Handling Missing Values

- **Model Training:** For CART, the DecisionTreeClassifier of Scikit-learn was employed, whose most important parameters are criterion, max depth, and minimum sample split, optimized via GridSearchCV, while pruning improved generalization. Naïve Bayes, implemented using CategoricalNB, calculated prior probabilities from class distributions and used feature scaling to ensure compatibility with probabilistic assumptions. The Random Forest model, with 200 estimators, was optimized by tuning parameters such as depth and split size, and feature importance rankings highlighted the significance of ECG and blood pressure metrics. CNNs were built with input layers for 187 features, followed by convolutional layers with 32 and 64 filters, max-pooling layers, dropout regularization 0.3, and a dense softmax layer. The CNN is trained using the Adam optimizer with a learning rate set to 0.001 and binary cross-entropy loss.

Model: "auto_encoder"

Layer (type)	Output Shape	Param #
sequential (Sequential)	(None, 1, 32)	63,264
sequential_1 (Sequential)	?	0 (unbuilt)

Total params: 63,264 (247.12 KB)
Trainable params: 62,688 (244.88 KB)
Non-trainable params: 576 (2.25 KB)

Fig. 11: Model Training

- **Model Evaluation:** The performance metrics used to evaluate the models are accuracy, precision, recall, F1-score, and confusion matrices. CART showed a very good accuracy of 92.77, whereas precision and recall were moderate; hence, it was considered a good baseline model. However, Random Forest outperformed CART with an accuracy of 97.69, where the precision was well-balanced at 91.16 and recall at 87.17, hence effectively using ensemble learning for robust feature interactions. While CNN had a good potential for capturing deep patterns in ECG signals, its performance was bounded due to class imbalance issues; thus, CNN achieved 46 in overall accuracy. Though the recall for CNN in anomaly detection was very high, with 87, the precision was much lower, suggesting a high number of false positives. This demonstrates that for cardiovascular anomaly detection, both machine learning, typically Random Forest, and deep learning methods, in this case CNN, are very relevant.

IV. Comparison and Analysis of Models

In this work, four models were implemented: CART, Naive Bayes, Random Forest, and CNN. Their performance was analyzed in detecting cardiovascular anomalies. Each model's strengths and weaknesses were assessed based on metrics such as accuracy, precision, recall, and F1-score, alongside considerations of computational efficiency and suitability for the medical domain. The evaluation provided insights into how different algorithmic approaches handle the complexities of cardiovascular data, including its imbalance and feature interdependencies.

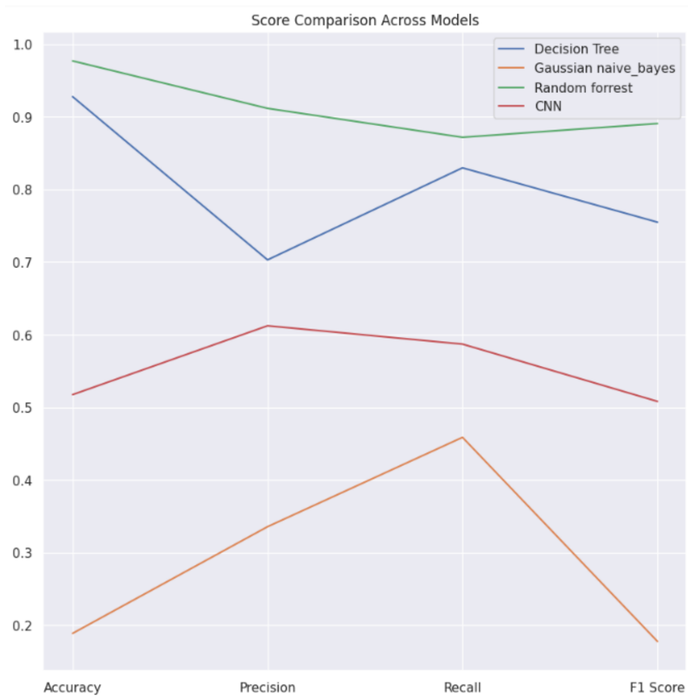


Fig. 12: Comparsion Line plot of Models

Among the models, Random Forest was the most accurate and robust, with a high accuracy of 97.69, a balanced precision, and recall. The ensemble-based approach not only reduced the risk of overfitting but also provided feature importance, which is valuable in medical diagnostics. On the other hand, Naive Bayes was computationally efficient, yet it was poor regarding feature correlations and imbalance of the dataset, hence returning accuracy and poor performance in the classification. CART was more balanced between interpretability and accuracy, though with higher values of false positives. Hence, pruning is inevitable to generalize well. At the same time, this example showed the ability of deep learning to extract raw ECG signals into meaningful hierarchical features using CNNs. While it achieved a strong recall for anomalies, its precision and overall accuracy were limited, largely due to the dataset imbalance and the computational complexity of training deep models.

Overall, the comparative study showed that classical machine learning models, such as Random Forest and CART, are reliable and interpretable, especially in medical fields where transparency is a major factor. Deep learning models like CNN, though posing certain challenges, hold great potential for improving anomaly detection via advanced optimization techniques: data augmentation, transfer learning, and weighted loss functions. This combination of classical and modern approaches constitutes a comprehensive toolkit in facing the challenges of cardiovascular anomaly detection.

A. *Advantages of the Implemented Models*

Each of the models implemented had its unique advantages for cardiovascular anomaly detection. CART is highly interpretable, which makes it a very good choice for medical applications where explainability is crucial. Its simple tree structure is easily understandable by clinicians in the decision-making process it comes up with. Naive Bayes is computationally efficient and works well on smaller datasets or preliminary analysis, returning quick results with very minimal computational overhead. It uses the strengths of multiple decision trees and proves to be robust to noisy data or unbalanced sets. In addition, with feature importance ranking, one has a deeper knowledge about which variable may be important-ECG Readings, blood pressure, and so on. On other hand, CNN works phenomenally while learning from the raw ECG signals where convolution layers automatically extract a hierarchy of features from small to higher level without explicit meddling. This makes CNNs especially apt at dealing with raw time-series data such as ECGs.

B. *Disadvantages of the Implemented Models*

Despite their strengths, the models also exhibited notable limitations. CART, while interpretable, is prone to overfitting in high-dimensional datasets, requiring pruning to improve generalization. Its moderate precision also indicates a susceptibility to false positives. Naive Bayes, with its assumption of feature independence, struggled to handle the dataset's correlated features, leading to poor overall performance. Random Forest, though robust and accurate, is computationally intensive, particularly with a large number of estimators, which may pose challenges in real-time applications. Additionally, the ensemble nature of Random Forest makes it less interpretable compared to single decision trees. CNN, while powerful, is highly sensitive to data imbalance, requiring techniques like weighted loss functions or data augmentation to mitigate its tendency for false positives. Moreover, CNNs are computationally expensive, demanding significant resources for training and requiring large, balanced datasets for optimal performance.

Each model has its strengths and limitations, and their combined use or further optimization could address the challenges posed by cardiovascular anomaly detection tasks.

V. Conclusion

The evaluation of different models in the detection of cardiovascular anomalies, with their strengths and weaknesses, provided a clear view on their applicability. Among these, the Random Forest model was the most reliable model, which provided very good accuracy and robustness through handling complex feature interactions, along with feature importance ranking. Though somewhat prone to overfitting, CART presents a pretty good balance between performance and interpretability, making it ideal for scenarios where explainability becomes crucial. On the contrary, Naive Bayes, though efficient and speedy in computation, falters on correlated features and/or unbalanced data, which greatly compromise its predictive accuracy. CNNs showed that there are meaningful patterns to be learned from raw ECG signals and used deep learning for anomaly detection, while still being computationally intensive, besides facing the challenge of imbalanced datasets. Handling challenges with appropriate techniques such as data augmentation, transfer learning, and feature selection would boost the performance of these models further. The key focus of this study is on the importance of including both traditional machine learning models and modern deep learning approaches to construct robust and interpretable solutions for medical diagnostics.

	Model	Accuracy	Precision	Recall	f1 score
0	Decision Tree	0.927736	0.703045	0.829737	0.754944
1	Gaussian naive_bayes	0.188882	0.335737	0.458747	0.177891
2	Random forrest	0.976932	0.911603	0.871762	0.890700
3	CNN	0.517508	0.612312	0.587083	0.508229

Fig. 13: Summary of Models

VI. Future Work

Future work can emphasize the improvement of cardiovascular anomaly detection by reducing data imbalance using weighted loss functions, synthetic oversampling, or data augmentation techniques. The use of pre-trained CNNs through transfer learning may boost performance with higher computational efficiency. Hybrid ensembling methods that incorporate more traditional models, such as Random Forest, with deeper learning approaches may result in higher accuracy and robustness. Further efforts could also be placed on optimizing models for real-time inference, enabling integration into clinical workflows. Expanding and diversifying the dataset to include a broader range of patient profiles would help improve the models' generalization and applicability in real-world medical settings.

References

1) Mohammad Kachuee, Shayan Fazeli, Majid Sarrafzadeh, "ECG Heart-beat Classification: A Deep Transferable Representation," Magnetism, University of California, Los Angeles (UCLA).
2) Li, H.; Boulanger, P. A Survey of Heart Anomaly Detection Using Ambulatory Electrocardiogram (ECG). Sensors 2020, 20, 1461. *https://doi.org/10.3390/s20051461*

3) Petrovic A, Jovanovic L, Venkatachalam K, Zivkovic M, Bacanin N, Budimirovic N. “Anomaly detection in electrocardiogram signals using metaheuristic optimized time-series classification with attention incorporated models.” *International Journal of Hybrid Intelligent Systems*. 2024;20(2):159-183.
doi:10.3233/HIS-240004
4) Z. Liu, W. Liu, Z. Gu and F. Yang, “Efficient Anomaly Detection Algorithm for Heart Sound Signal,” in *IEEE Access*, vol. 12, pp. 139225-139236, 2024.
doi:10.1109/ACCESS.2024.3465540.
5) M. Zhang, A. Raghunathan and N. K. Jha, “MedMon: Securing Medical Devices Through Wireless Monitoring and Anomaly Detection,” in *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 6, pp. 871-881, Dec. 2013.
doi:10.1109/TBCAS.2013.2245664.