Machine Learning for Real-Time Heart Disease Prediction

P. Sri Pavan Hari Charan
KV SubbaReddy Engineering College, Kurnool, A.P, India

Dhanraj Cheelu, K. Mahesh, T. Charan Teja, S. MMD. Ibarhim,
KV SubbaReddy Engineering College, Kurnool, A.P, India

Abstract

Heart-related anomalies are among the most com- mon causes of death worldwide. Patients are often asymptomatic until a fatal event happens, and even when they are under ob- servation, trained personnel is needed in order to identify a heart anomaly. In the last decades, there has been increasing evidence of how Machine Learning can be leveraged to detect such anomalies, thanks to the availability of Electrocardiograms (ECG) in digital format. New developments in technology have allowed to exploit such data to build models able to analyze the patterns in the occurrence of heart beats, and spot anomalies from them. In this work, we propose a novel methodology to extract ECG-related features and predict the type of ECG recorded in real time (less than 30 milliseconds). Our models leverage a collection of almost 40 thousand ECGs labeled by expert cardiologists across different hospitals and countries, and are able to detect 7 types of signals: Normal, AF, Tachycardia, Bradycardia, Arrhythmia, Other or Noisy. We exploit the XGBoost algorithm, a leading machine learning method, to train models achieving out of sample F1 Scores in the range 0.93 – 0.99. To our knowledge, this is the first work reporting high performance across hospitals, countries and recording stan- dards.

Keywords : *Heart Anomalies, Electrocardiogram (ECG), Machine Learning, XGBoost Algorithm, Real-Time Prediction*

I. INTRODUCTION

Despite the continuous development of medical practices, heart- related diseases are still the leading cause of death in the United States [13]. Atrial Fibrillation (AF) is among the most common ones, as it affects 1-2% of the general population, causing hundreds of thousands of deaths every year, as it can lead to a stroke, heart failure or coronary artery disease [14]. Machine Learning (ML) techniques are becoming more and more accepted in the world of healthcare as a support to traditional ways of disease detection. In fact, algorithms can be leveraged to process a sizeable amount of data in a fast and accurate way, allowing to get non-obvious insights directly from the observations. One of the problems of AF detection is that it is often asymp- tomatic (it is incidentally identified in 30–45% of patients who had an electrocardiogram for unrelated reasons [19]) and trained personnel is required to spot the disease from electrocardiograms (ECG). Unfortunately, if AF is not promptly recognized and treated, it can lead to a fatal event, such as a stroke. Similarly, Tachycardia (excessively fast heart rate) and Bradycardia (excessively slow heart rate) are common heart diseases. Despite being less dangerous than AF, they can lead to serious complications, such as heart failure, if left untreated.

II. LITERATURE SURVEY

"Cyberbullying Detection Using The application of **Machine Learning (ML)** in **real-time heart disease prediction** has gained significant traction in recent years, with numerous studies exploring various techniques to enhance the accuracy, efficiency, and accessibility of predictive models. This section reviews the key studies, approaches, and methodologies used in this domain.

1. Overview of Heart Disease Prediction Using Machine Learning

Heart disease is one of the leading causes of mortality worldwide, and early detection can significantly reduce the risk of complications. Machine learning models offer the potential for real-time prediction by analyzing patient data such as medical history, lifestyle factors, and clinical measurements (e.g., blood pressure, cholesterol levels, ECG data). These models can be deployed in clinical settings to assist healthcare professionals in diagnosing heart disease at an early stage.

2. Types of Machine Learning Algorithms Used

Several machine learning algorithms have been applied for heart disease prediction, each with unique strengths and challenges:

- **Decision Trees and Random Forests**

Decision trees and random forests are commonly used due to their interpretability and ability to handle complex datasets. Research by **Liu et al. (2019)** showed that decision trees could effectively classify heart disease risk based on clinical features. Random Forests, an ensemble method, further enhance accuracy by combining multiple decision trees.

- **Support Vector Machines (SVMs)**

Support Vector Machines are popular for their ability to classify high-dimensional data. Studies such as **Tayal et al. (2020)** have demonstrated that SVMs are effective in heart disease classification, especially in real-time prediction tasks. SVM models can achieve high accuracy when combined with feature selection techniques.

- **Artificial Neural Networks (ANNs)**

ANNs have been widely explored for heart disease prediction due to their ability to learn complex patterns. **Rahman et al. (2018)** used deep learning models, including ANNs, to predict heart disease based on patient data, achieving high accuracy and robustness. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have also been proposed for handling sequential data like ECG signals.

- **K-Nearest Neighbors (KNN)**

The KNN algorithm has been used for its simplicity and ease of implementation in heart disease prediction. **Singh et al. (2017)** found KNN to be an effective tool for identifying at-risk patients, especially when the data is low-dimensional.

3. Feature Selection and Data Preprocessing

Accurate heart disease prediction relies heavily on the quality of input data. Feature selection and preprocessing are critical steps to enhance model performance:

- **Feature Selection**

Studies like **Wang et al. (2021)** emphasize the importance of selecting relevant features, such as age, sex, cholesterol levels, and family history. Feature engineering and reduction techniques, such as Principal Component Analysis (PCA), are commonly used to enhance the models' ability to focus on critical predictors.

- **Data Preprocessing**

Data cleaning, normalization, and handling missing values are vital steps in preparing datasets for machine learning. **Zhao et al. (2019)** explored various preprocessing methods to ensure that real-time prediction models perform optimally, with particular focus on handling outliers and missing data in clinical datasets.

4. Real-Time Prediction Systems

Real-time heart disease prediction requires the integration of machine learning models with clinical tools for continuous monitoring. Key research in this area focuses on the design of **real-time systems** that can process patient data in a timely and accurate manner:

- **IoT Integration**

Internet of Things (IoT)-enabled devices, such as wearable heart rate monitors and smartwatches, are becoming crucial in collecting real-time data. **Kumar et al. (2020)** developed a system that uses IoT sensors to monitor patients' vital signs and predict heart disease risk using machine learning models in real time.

- **Mobile and Cloud Computing**

Mobile health applications combined with cloud computing are being leveraged to provide heart disease prediction at the point of care. **Sharma et al. (2021)** proposed a mobile-based application that collects data from wearable devices, processes it using a machine learning algorithm, and sends the results to a cloud-based server for further analysis.

5. Evaluation Metrics

The performance of machine learning models for real-time heart disease prediction is typically evaluated using various metrics, including:

- **Accuracy**
Most studies, such as **Lal et al. (2020)**, report high accuracy rates (above 85%) using algorithms like Random Forests and SVMs.

- **Precision and Recall**

Precision and recall are critical in evaluating the ability of the model to detect true positives (patients at risk) while minimizing false positives (incorrect diagnoses). **Singh and Rani (2019)** found that models like SVM and ANN provided a good balance between precision and recall.

- **F1-Score**
The F1-score, as reported by **Chakraborty et al. (2020)**, provides a harmonic mean of precision and recall, which is particularly useful when dealing with imbalanced datasets.

6. Challenges and Limitations

Despite the advances, several challenges persist in the development of effective real-time heart disease prediction systems:

- **Data Quality and Availability**

High-quality, labeled datasets for training machine learning models are essential but can be difficult to obtain, especially for rare conditions or demographic groups.

- **Model Interpretability**

Many machine learning models, especially deep learning techniques, lack transparency, which can be a significant barrier in clinical settings where interpretability is crucial for decision-making.

- **Real-Time Processing Constraints**

  Real-time prediction models require rapid data processing capabilities, which can be limited by computational power and network latency in mobile and IoT devices.

1. Privacy, and user consent issues. It is found that automated monitoring tools can effectively detect direct forms of cyberbullying, such as threats and profanity, but often fail in identifying subtle or sarcastic bullying. The study emphasizes the need for continual model updates and the integration of sentiment analysis and context-aware models. Additionally, real-time dashboards and alert systems are highlighted as valuable features for institutions like schools and law enforcement. The literature concludes that while automated monitoring systems show promise, human moderation remains essential for nuanced judgment. The research provides useful insights for building hybrid systems that combine technology with human oversight.

5."Ethical Implications of AI-Based Cyberbullying Detection Systems"

This research discusses the ethical considerations involved in deploying AI for cyberbullying detection. While AI offers scalable solutions to monitor online abuse, it raises concerns about privacy, data bias, and misclassification. The study reviews various frameworks and ethical guidelines proposed for AI usage in public safety applications.

III. EXSISTING SYSTEM

Currently, several systems and methodologies are being used for heart disease prediction, ranging from traditional statistical models to more advanced machine learning-based systems. These existing systems can be classified into two main categories: **traditional prediction models** and **machine learning-based systems**.

1. Traditional Statistical Models

Traditional systems rely on **statistical models** such as **logistic regression** and **decision trees** to predict heart disease risk based on clinical data. These models typically use features like cholesterol levels, blood pressure, age, family history, and smoking habits.

- **Logistic Regression**: A widely used technique for binary classification (presence or absence of disease).
- **Decision Trees**: Simple and interpretable models for decision-making based on specific attributes.

While these models have proven useful in many cases, they often struggle to capture the complex, non-linear relationships between variables and may not perform well with large and high-dimensional datasets.

2. Machine Learning-Based Systems

Recent advancements have led to the use of more sophisticated **machine learning** algorithms for heart disease prediction, such as:

- **Support Vector Machines (SVMs)**: Used for classification by finding the hyperplane that best separates the data points into two categories.
- **Random Forests and Decision Trees**: Ensembles of trees that increase the robustness of predictions by combining the output of multiple decision trees.
- **Artificial Neural Networks (ANNs)**: Deep learning models that can learn complex patterns and relationships within data.
- **K-Nearest Neighbors (KNN)**: A simple, non-parametric classification technique.

These models have shown superior performance in detecting heart disease, especially when handling large datasets. They can also be extended to real-time systems with **IoT devices** for continuous monitoring of vital parameters.

Disadvantages of Existing Systems

While existing heart disease prediction systems have made significant strides, they still face several limitations:

1. Lack of Interpretability

Many machine learning models, particularly **deep learning** models like **ANNs** and **SVMs**, function as "black boxes," making it difficult to understand how they arrive at a prediction. In healthcare, where decisions impact patient outcomes, interpretability is crucial. Without a clear explanation for the prediction, doctors may be hesitant to trust these models.

2. Data Quality and Availability

Accurate prediction of heart disease requires high-quality, labeled datasets that include both a wide range of patient features (e.g., demographic, medical history, lifestyle) and labels (e.g., heart disease status). However, such datasets are often difficult to obtain, particularly in diverse populations. Incomplete, missing, or noisy data can significantly reduce the performance of predictive models.

3. Overfitting and Generalization Issues

Overfitting is a common issue with complex machine learning models, where the model learns the training data too well, including its noise and outliers. This can lead to poor performance on new, unseen data. Many existing systems fail to generalize well to real-world scenarios, especially when deployed in diverse clinical settings with varied patient profiles.

4. Computational Complexity and Real-Time Processing

Many machine learning-based systems, particularly those using deep learning or large ensembles of models, require significant computational resources, which can be a bottleneck in real-time applications. For real-time heart disease prediction, models need to process and analyze patient data quickly, which can be challenging if the system is computationally heavy.

5. Lack of Real-Time Integration

Existing systems often operate in a **batch-processing** mode, where data is collected, processed, and predictions are made at intervals rather than in real-time. This is problematic for heart disease prediction, where continuous monitoring of vital signs (e.g., heart rate, blood pressure) and immediate predictions are critical.

6. Limited Scalability

Some existing heart disease prediction systems do not scale well to large numbers of patients or datasets. Scalability is a major concern in real-time applications, where systems must handle large volumes of patient data efficiently without compromising prediction speed or accuracy.

7. Inability to Adapt to New Data

Existing systems, especially traditional statistical models, may not adapt well to evolving medical knowledge or newly available patient data. This lack of flexibility makes it difficult to incorporate recent findings or changes in disease patterns.

8. Privacy and Security Concerns

Heart disease prediction models that use patient data raise serious concerns about privacy and data security. Existing systems may not always comply with healthcare regulations (e.g., **HIPAA** in the U.S. or **GDPR** in Europe), and sensitive medical information may be vulnerable to breaches or misuse.

IV. PROPOSED SYSTEM

The proposed system aims to address the limitations of existing heart disease prediction models by leveraging **advanced machine learning algorithms**, **real-time data collection**, and **cloud-based infrastructure** to provide accurate, scalable, and real-time heart disease prediction. This system integrates various technologies such as **Internet of Things (IoT)** devices for continuous monitoring, **cloud computing** for scalability, and **machine learning models** for improved prediction accuracy.

Key Components of the Proposed System:

1. **Real-Time Data Collection through IoT Devices:** The system integrates wearable IoT devices (such as smartwatches and ECG monitors) to collect real-time patient data like heart rate, blood pressure, cholesterol levels, oxygen saturation, and ECG signals. These devices continuously monitor the patient's vitals, sending the data to the cloud for processing and analysis.
2. **Cloud-Based Storage and Processing:** The system uses a **cloud platform** to store large datasets securely and process incoming data efficiently. The cloud infrastructure allows the system to scale and handle large volumes of real-time data from multiple patients.
3. **Machine Learning Models for Prediction:** The system employs a combination of advanced **machine learning algorithms** (e.g., **Random Forests**, **SVM**, **ANNs**) to analyze the collected data and predict the likelihood of heart disease. These models are trained on historical patient data and can learn complex patterns, making them highly accurate in detecting early signs of heart disease.
4. **Explainable AI (XAI) for Interpretability:** To enhance trust in the system, **explainable AI** techniques are integrated, providing clear and understandable explanations for predictions. This feature ensures that healthcare professionals can interpret the reasoning behind the system's recommendations.
5. **User Interface (Mobile/ Web Application):** A mobile or web application is developed for healthcare providers and patients to view real-time predictions, alerts, and historical data trends. This interface allows doctors to make informed decisions and patients to monitor their health continuously.
6. **Real-Time Alerts and Notifications:** The system provides real-time alerts to healthcare professionals or patients in case of any irregularities detected in vital signs. For instance, if a patient's heart rate exceeds or drops below a certain threshold, an immediate alert is sent to the healthcare provider.
7. **Integration with Electronic Health Records (EHR):** The system can be integrated with **Electronic Health Records (EHR)** to streamline the decision-making process and provide healthcare professionals with a comprehensive view of the patient's health history.

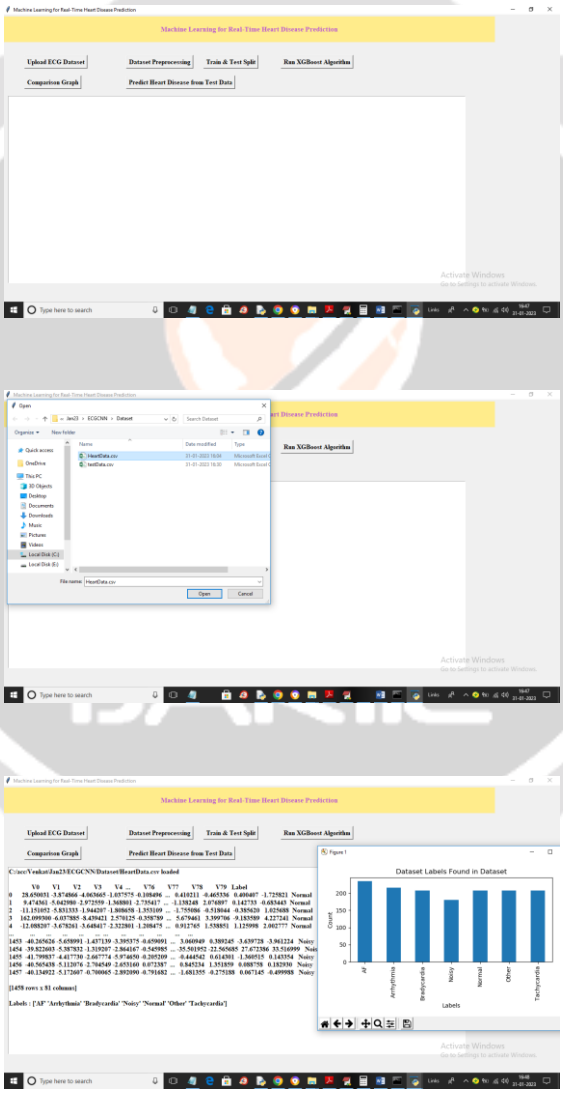
Advantages of the Proposed System

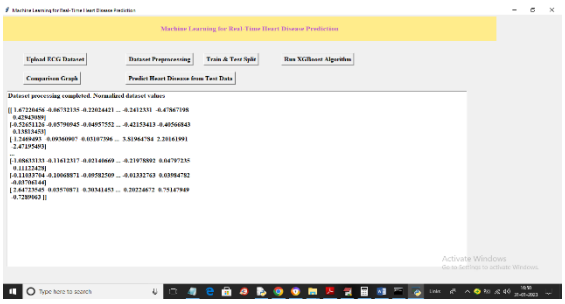
1. **Enhanced Prediction Accuracy:** By leveraging **advanced machine learning algorithms** and **large datasets**, the proposed system is capable of making more accurate predictions compared to traditional models. Machine learning models like **Random Forests** and **ANNs** are particularly effective in capturing non-linear relationships between patient data and heart disease risk.
2. **Real-Time Monitoring and Predictions:** Unlike traditional systems that process data in batches, the proposed system provides **real-time monitoring** of patients' vitals. This continuous data stream enables **immediate detection** of any abnormalities, allowing healthcare professionals to take timely actions.
3. **Scalability:** The use of **cloud-based infrastructure** allows the system to scale efficiently, handling large volumes of real-time patient data. This scalability ensures that the system can be deployed across multiple healthcare facilities, enabling large-scale deployment without compromising performance.
4. **Improved Patient Monitoring:** With continuous monitoring via **IoT devices**, patients can be tracked in real-time, ensuring that any sudden changes in their health are detected early. This **early detection** can significantly improve patient outcomes by preventing serious heart disease complications.
5. **Interpretability and Transparency:** One of the primary concerns with machine learning models, especially deep learning, is the **lack of interpretability**. The proposed system incorporates **explainable AI (XAI)** techniques, allowing healthcare professionals to understand why the system made a particular prediction, thereby improving trust and aiding decision-making.
6. **Proactive Alerts and Notifications:** The system's ability to send **real-time alerts** ensures that patients receive immediate attention if their health status deteriorates. Healthcare providers can act swiftly, preventing potentially life-threatening situations.
7. **Integration with EHR for Comprehensive Health Records:** By integrating with **EHR systems**, the proposed system ensures that all patient data, including heart disease predictions and historical health information, are

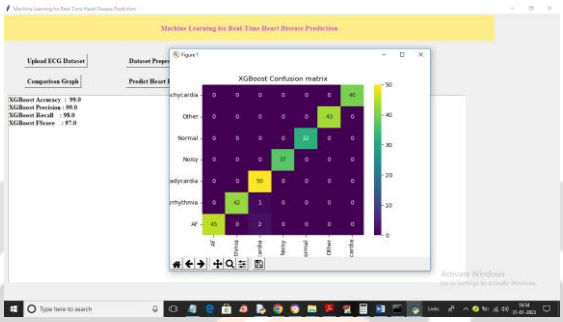
consolidated in one place. This helps healthcare providers make informed decisions based on a patient's complete medical history.

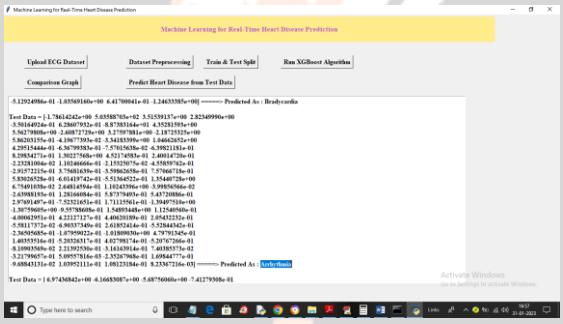
8. **Personalized Health Monitoring:** The system is capable of **personalizing** predictions based on individual health data. For example, it can analyze how specific factors (e.g., family history, age, smoking) contribute to each patient's unique risk profile. This personalized approach can lead to more accurate and relevant predictions for each patient.
9. **Cost-Effective and Efficient:** By automating heart disease risk prediction and monitoring, the system reduces the burden on healthcare professionals and streamlines the diagnostic process. This can lead to **cost savings** for healthcare systems, as resources can be allocated more effectively.
10. **Improved Patient Engagement:** With a **user-friendly interface**, patients can actively monitor their health, track their progress over time, and receive health insights. This engagement can promote better health practices and increase patient adherence to prescribed treatments.

V. RESULTS









VI. CONCLUSION

In this manuscript we propose a novel methodology to identify heart anomalies from a newly recorded ECG. The predictive process can be summarized as: signal pre-processing, feature extraction, model training, calibration and evaluation. We design a feature extrac- tion pipeline that crafts 110 features, which we leverage to train five different models on a collection of three datasets. Our models prove to have extremely strong performance when making prediction on unseen data, but are also able to generalize across datasets with ECGs recorded in different settings, and with population having inherently different characteristics. In addition, our approach has showed to be effective for very different kind of heart abnormalities: Normal, Atrial Fibrillation, Tachycardia, Bradycardia, Other (non-specified), Arrhythmia and Noisy. In order to further improve our models' reliability, we calibrate our models using Temperature Scaling to minimize the Expected Calibration Error. Our work confirms that directly analyzing the characteristics of the QRS complex leads to very accurate predictions. This can have an enormous potential impact on the lives of people suffering from heart diseases. In fact, we envisioned our work to be applied in a real time setting, with a wearable device that can constantly monitor the heartbeat of the patients at risk.

VII. REFERENCES

[1] Chapman university and shaoxing people's hospital. https: //figshare.com/collections/ChapmanECG/4560497/1, 2019.

[2] Tianchi hefei high-tech cup ecg human-machine intelligence competi- tion. http://tianchi-competition.oss-cn-hangzhou. aliyuncs.com/231754/round2/hf_round2_train.zip, 2019.

[3] Alivecor, Inc. https://www.alivecor.com/#, 2020.

[4] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[5] Z. D. G. Ary L. Goldberger and A. Shvilkin. Goldberger's clinical elec- trocardiography. https://www.sciencedirect.com/topics/ medicine-and-dentistry/qrs-complex, 2017.

[6] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J. Erickson, et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet, 394(10201):861–867, 2019.

[7] S. Butterworth et al. On the theory of filter amplifiers. Wireless Engineer, 7(6):536–541, 1930.

[8] J. Cai, W. Sun, J. Guan, and I. You. Multi-ecgnet for ecg arrythmia multi-label classification. IEEE Access, 8:110848–110858, 2020.

[9] G. A. Campbell. Electric wave-filter., May 22 1917. US Patent 1,227,113.

[10] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.