

Examining the Impact of Data Imbalance on the Effectiveness of the Proposed Algorithm for Real-Time Prediction of Heart Disease and Suggesting Solutions

Pragathi, Sneha Ghode, Nisha Hebbar, Bhoomika Surendra Naik, Professor Dr. Lokesh M R
Dept. of Information Science and Engineering AJ Institute of Engineering and Technology Mangalore, India

Abstract- The main goal of this review paper is to describe the impact that data imbalance has on the Prophet algorithm's ability to accurately predict heart disease in real time and offer solutions for these effects. This case study of a mobile health illness management software includes three modules: User, Admin, and Doctor. ECG data is available to registered users, but they must also upload it in csv format so that the Prophet program, which generates educational reports, can further analyze it. The credibility of the predictions may be impacted by this biased data or noise, which could impair the model's performance and lead to biases in its output. This review highlights some issues with the problem of imbalanced data, makes an effort to gather data and knowledge from the literature that is currently available, and offers some tactics that could be helpful in resolving such problems, including algorithm modification, data resampling, and synthetic data. The article concludes by discussing the application's potential to improve heart disease early detection and streamline interactions between medical professionals and patients. Because it deals with healthcare quality, this review first offers a framework for future research on mobile health technologies and emphasizes the idea of addressing data imbalance in data-driven models.

Index Terms- Data Imbalance, Heart Disease Prediction, Mobile Health Technology, Prophet Algorithm, Real-Time Monitoring

I. INTRODUCTION

The rapid development of mobile health technology has revolutionized healthcare delivery, particularly in the areas of prognosis assessment, illness prevention, and management. The creation of real-time cardiovascular disease prediction is one noteworthy example; it has gained widespread recognition as a crucial tool for enhancing patient outcomes via prompt treatment [1]. The complex algorithms that underpin these predictive models analyze patient data to generate insightful conclusions. Among these, the Prophet algorithm has shown promise in predicting health events; however, its efficacy is frequently compromised by the prevalent issue of data imbalance in healthcare datasets [2].

Data imbalance occurs when specific illnesses, such as heart disease, are underrepresented within a dataset compared to healthier examples [3]. This disparity could eventually lead to poor clinical outcomes since biased prediction models would be unable to accurately identify patients who are at risk.

In the case of cardiac disease prediction, when prompt and precise diagnosis is crucial, the consequences of data imbalance are especially worrisome [4]. Section II goes deeper on this very important topic and emphasizes the need

for practical solutions to mitigate the effect of data imbalance on predictive algorithms.

This review is divided into several critical sections. Focusing on two key datasets, namely the ECG Signal Data and the Symptoms Dataset, the first section presents a comprehensive review of the input data types employed in heart disease prediction [5]. These datasets are necessary to understand the clinical history of the patient and the electrical activity of the heart, respectively. We may be able to comprehend the nature of the issues related to data imbalance and how it affects model performance by taking a closer look at the traits and attributes of these datasets as covered in Section III. Section IV explains the significance of explainable AI technology in cardiac care after discussing input data [6]. This section stresses the importance of model interpretability in clinical decision-making, patient involvement, and regulatory compliance. Since the healthcare industry is depending more and more on AI- driven predictions, it is critical to comprehend their motivations in order to foster trust between patients and healthcare providers [7]. The issue of data imbalance in heart disease prediction is covered in Section V, which follows. This issue includes biases and reduced sensitivity in training algorithms when using unbalanced datasets [8]. Particularly for healthcare applications, this

discussion highlights the importance of having robust evaluation metrics that provide an equitable representation of model performance across classes. Understanding the limitations of traditional evaluation methods in imbalanced datasets requires

familiarity with the ideas covered below.

Section VI offers a number of techniques to enhance Prophet algorithm performance in order to address the issue of data imbalance [9]. All of these strategies—from ensemble approaches to cost-sensitive learning and advanced resampling techniques—may result in predictive models that are more accurate and balanced. This section summarizes the body of existing literature and offers practical solutions to the issues raised above in an effort to provide a comprehensive framework for further research on this topic.

The therapeutic implications of integrating explainable AI technology into cardiac care are covered in Section VII, the review's final section. It shows the potential for personalized treatment regimens, real-time monitoring, and better patient outcomes [10]. By concentrating on the problems of data imbalance and model interpretability, healthcare providers can use AI-driven solutions to improve the quality of care given to patients with heart disease. All of the concepts that come out of this paper are combined in this conclusion, which emphasizes the necessity of controlling data imbalance in predictive analytics.

II. MOTIVATION

Since cardiovascular disease is still the major cause of morbidity and death globally, there is an evident need for useful predictive techniques to assist in the early detection and treatment of cardiovascular illness. The use of mobile health (mHealth) technologies in healthcare settings has opened up clinically promising new avenues for real-time patient data collection and analysis [14][15]. However, data imbalance—a common problem with health care data—typically restricts the performance of predictive algorithms (such as the Prophet model) [16].

When certain labels, like heart diseases (HD), are underrepresented in the observed dataset compared to other labels, this is known as data imbalance. Models may not be able to accurately predict an individual's high risk of developing heart disease due to this skew distribution, which results in inaccurate predictions. Therefore, these projections may mislead medical professionals, resulting in poor patient-centered care and delayed diagnosis [17]. However, addressing data imbalance is crucial to improving a predictive model's validity and reliability in these circumstances.

A unique opportunity to improve the predictive capabilities of heart disease models is provided by the use of a mobile health application that permits the uploading and analysis of electrocardiogram (ECG) data. Using user-generated data, such an application can highlight the significance of high-quality data and offer extremely useful insights. However, managing dataset imbalance and noise effectively is essential to the success of this kind of effort.

This review's goal is to thoroughly investigate the connection between data imbalance and the Prophet algorithm's performance when it comes to heart disease prediction. The paper offers a thorough framework for future research in the field by synthesizing the literature from earlier studies on data imbalance mitigation strategies. Ultimately, another goal is to help optimize predictive analytics in medicine, which is made possible by mHealth solutions that provide actionable and reasonably accurate information to help improve patient outcomes.

Nature of Data

For the purposes of this review paper, the input data consists of two primary datasets: ECG Signal Data and Symptoms Dataset. Both datasets are structured in CSV format, allowing for easy manipulation and analysis.

ECG Signal Data

The ECG signal dataset is crucial for analysing the electrical activity of the heart. Each record in the CSV file represents a single ECG reading.



Fig 1: Different types of ECG leads used in recording electrical activity of the heart. Each lead provides a different view of the heart's electrical activity, which can help in diagnosing various conditions.

Raw ECG Signal

The raw ECG signal is recorded as a time series where each sample represents a voltage measurement at a specific point in time. If the ECG signal is sampled at 250 Hz, the time interval (T) between samples can be calculated as in (1).
$$T = 1/f_s = 1/250\text{Hz} = 0.004 \text{ seconds} = 4\text{ms} \quad (1)$$

Thus, the raw ECG signal is represented as a time series $V(t)$, where t represents time in seconds.

Sample ECG Signal

A typical ECG waveform includes the following distinct features:

- **P wave:** Depolarization of the atria.
- **QRS complex:** Ventricular depolarization (with the R-peak being the most prominent feature).
- **T wave:** Ventricular repolarization.

These features correspond to different phases of the cardiac cycle.

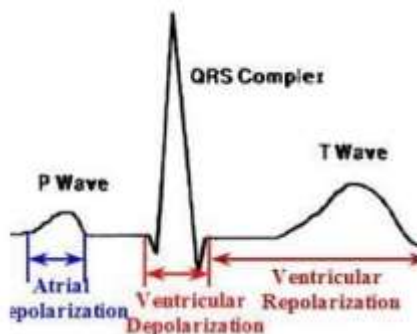


Figure 2: Sample ECG waveform illustrating the electrical activity of the heart over time. The waveform typically includes distinct features such as P waves, QRS complexes, and T waves, which correspond to different phases of the cardiac cycle.

Heart Rate Calculation

Finding the R-peaks in the QRS complexes over a predetermined amount of time allows one to determine the heart rate (HR) from the ECG signal. If N is the number of R-peaks found in a time interval T (in seconds), as in (2), the heart rate in beats per minute (BPM) can be calculated as follows.

$$HR = (N/T) \times 60 \quad (2)$$

Arrhythmia Indicators

The ECG signal can be used to identify arrhythmias, which are irregular cardiac rhythms. Typical arrhythmias consist of:

- **Atrial fibrillation (AF):** Irregular, rapid heart rate due to chaotic atrial activity.
- **Ventricular tachycardia (VT):** Abnormally fast heart rhythm originating in the ventricles. To identify arrhythmias, algorithms analyse the RR intervals (time between successive R-peaks):
- **Regularity:** Check for consistent RR intervals (e.g., irregular intervals may indicate AF).
- **Rate:** An elevated heart rate (e.g., >100 BPM) may indicate VT or supraventricular tachycardia (SVT).
- **Morphology:** Analyse QRS complex shapes to detect anomalies.

Arrhythmia indicators are generated by these algorithms and flag specific conditions for further analysis.

Symptoms Dataset

A. Clinical observations and patient self-reported symptoms are included in the symptom's dataset, which provides contextual data that enhances predictive modelling. The ECG Signal Data and the Symptoms Dataset are the two enormous datasets that make up the input data for this review paper. Using a variety of features, such as patient identification, timestamps, lead types, raw voltage measurements, heart rates, and arrhythmia indicators, ECG signal data records the intricate electrical activity of the heart. This dataset is essential for diagnosing heart conditions and determining how well the heart is functioning. A more comprehensive picture of the patient's cardiovascular health is made possible by the Symptoms Dataset, which supplements this by offering insightful clinical context in the form of self-reported symptoms and medical history. When combined, these datasets improve clinical decision-making, facilitate predictive modelling, and enable a more thorough examination of heart disease.

Enhancing Cardiac Care With Explainable AI Technologies

The integration of explainable artificial intelligence (AI) technologies into cardiac care represents a transformative approach to diagnosing and predicting heart diseases. As healthcare increasingly relies on data-driven decisions, the importance of understanding the underlying mechanisms of AI algorithms becomes paramount. This is especially true for predictive algorithms such as the Prophet algorithm, which may be used to predict cardiac disease in real time and is specifically intended for time series forecasting.

Explainability's Significance in Heart Care

In artificial intelligence, explainability is the extent to which a model's internal workings are human-understandable. This is important when it comes to cardiac care for a number of reasons:

- **Clinical Decision-Making:** To make well-informed treatment recommendations, doctors must comprehend the reasoning behind forecasts. An explainable model increases confidence in the AI system by revealing which factors have the biggest effects on the outcomes that are predicted.
- **Patient Engagement:** Patients are more likely to adhere to treatment plans when they understand the reasoning behind their diagnosis or prognosis. Better communication between patients and healthcare professionals is facilitated by explainable AI, which raises patient satisfaction and compliance.
- **Regulatory Compliance:** Explainable AI technologies can assist healthcare organizations in meeting regulatory bodies' increasing emphasis on the need for transparency

in AI applications, thereby guaranteeing ethical standards in patient care.

Challenges of Data Imbalance in Heart Disease Prediction

Data imbalance is a significant obstacle when using the Prophet algorithm to predict heart disease. Heart disease cases (the positive class) are substantially less common than healthy cases (the negative class) in many datasets. This disparity may result in:

- **Bias in Predictions:** Due to the model's bias towards predicting the majority class, a sizable number of false negatives may result, potentially missing actual cases of heart disease.
- **Reduced Sensitivity:** There could be major clinical repercussions if a model trained on unbalanced data shows decreased sensitivity, failing to detect patients at risk for heart disease.

Implementing techniques that improve the Prophet algorithm's performance while preserving its explainability is crucial to overcoming these obstacles.

Strategies to Address Data Imbalance

Several strategies can be employed to mitigate the effects of data imbalance on the Prophet algorithm's performance:

- **Resampling Techniques:** Implementing techniques such as Synthetic Minority By balancing the dataset and producing artificial examples of the minority class (heart disease cases), the Over-sampling Technique (SMOTE) can help the model learn from these examples more successfully.
- **Cost-Sensitive Learning:** Assigning higher costs to misclassifying the minority class can encourage the model to focus more on accurately predicting heart disease cases, thereby improving sensitivity without compromising overall accuracy.
- **Ensemble Methods:** Predictive performance can be improved by employing ensemble techniques or combining several models. For example, combining the Prophet algorithm with other models can offer a more thorough understanding of patient data, increasing explainability and prediction accuracy.

Model Comparison and Evaluation

A critical aspect of the research involves comparing the performance of the Prophet algorithm against traditional models like ARIMA and advanced deep learning techniques. This comparison is vital for:

- **Identifying Strengths and Weaknesses:** Researchers can determine which models work best in particular clinical situations by assessing how each model manages data imbalance and offers justifications for its predictions.

- **Understanding Predictive Factors:** Better clinical insights can be obtained by examining the Prophet algorithm's output in conjunction with that of other models to determine which variables have the greatest bearing on heart disease prediction.

Clinical Implications and Future Directions

The integration of explainable AI technologies in cardiac care has profound clinical implications:

- **Real-Time Monitoring:** By using the Prophet algorithm to predict heart disease in real time, medical professionals can better monitor their patients, enabling prompt interventions and possibly averting negative outcomes.
- **Personalized Treatment Plans:** Understanding the factors that influence predictions enables clinicians to tailor treatment plans based on individual patient profiles, leading to better outcomes.
- **Research and Development:** Future research should focus on expanding datasets to include diverse populations and various vital signs beyond heart rate. This will enhance the generalizability of the models, ensuring their applicability across different demographics and clinical settings.

Explainable AI technologies, particularly the Prophet algorithm, hold promise for enhancing the prediction of heart disease. These cutting-edge technologies can help healthcare providers make patient-centered, well-informed decisions by addressing data imbalance and placing a strong emphasis on model interpretability. Better clinical results and greater trust in AI-powered medical solutions are the outcomes of this, which pave the way for more effective and transparent cardiac care.

Datasets Available

Access to high-quality datasets is necessary for the training and evaluation of machine learning models, like the Prophet algorithm, for the prediction of heart disease. Numerous publicly available datasets offer a range of characteristics and features that can be utilized to improve explainability and predictive accuracy, as shown in Table 1.

Table 1: Numerous publicly accessible datasets are frequently used for research and prediction on heart disease.

Dataset Name	Source	Key Features	Instances	Class Distribution
UCI Machine Learning Repository: Heart Disease Dataset	UCI ML Repository	Age, sex, blood pressure, cholesterol, etc.	303	45.9% with heart disease, 54.1% without
Kaggle:	Kaggle	Age, sex,	299	6.8% with

Heart Failure Prediction Dataset		blood pressure, medical history		heart failure, 53.2% without
PhysioBank: MIMIC-III Critical Care Database	PhysioBank	Over 300 attributes (vital signs, lab results, medication information)	40,000+	Varies by specific task
National Institutes of Health: Framingham Heart Study Dataset	NIH	Over 100 attributes (demographic, clinical, lifestyle factors)	5,000+	Varies by specific task
Google Dataset Search: Heart Disease Datasets	Google Dataset Search	Varies by dataset	Varies	Varies by dataset

Challenges and Recommendations

Challenges In Data Imbalance

Limited Representation of Minority Class: One of the main obstacles to the prediction of heart disease is the minority class's underrepresentation in the majority of datasets. It's possible that the majority class, in this case, has biased models. Low sensitivity and more false negatives are the results of these biased models. A lack of appropriate data for the minority cases results in a very bad situation in terms of predictive accuracy because patients with heart diseases do not exhibit any symptoms until the disease is advanced.

Overfitting in Predictive Models: These are susceptible to overfitting on the majority class of unbalanced datasets. Even though forecast models can be made accurately, they are unable to generalize to new data at the same time, particularly when representative instances reflecting minority classes occur. In practical deployment and even perhaps under varying distributions of cases, it can end in a good training data-set performance model but a poorly performing model.

Difficulty in Model Evaluation: An important evaluation challenge is the performance of predictive models on unbalanced datasets. Conventional metrics, such as accuracy, can also be deceptive because a model may be highly accurate for the majority class but extremely inaccurate for the minority class. In order to provide a suitable assessment of the model's true performance across classes, this necessitates the use of more robust evaluation metrics, such as precision, recall, F1-score, and area under the ROC curve.

Recommendations for Improvement

Implementing Advanced Resampling Techniques: Addressing the problem of data imbalance requires advanced resampling techniques other than simple oversampling or under sampling. Advanced methods include SMOTE

(Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling), which synthesize examples of the minority class and enrich the training dataset for better model performance.

Utilizing Ensemble Learning Approaches: Ensemble learning methods, such as bagging and boosting, can be effective in improving model robustness against data imbalance. These techniques can leverage the strengths of each to enhance overall predictive performance by combining multiple models. For example, using a combination of the Prophet algorithm with decision trees or random forests can provide a more balanced approach to prediction.

Incorporating Cost-Sensitive Learning: This might go a long way to mitigate the issue of class imbalance with cost-sensitive learning by charging higher misclassification costs to the minority class, making it more sensitive to correctly predict instances of heart disease.

Continuous Monitoring and Updating of Models: Given the dynamic nature of healthcare data, it is important to have a system for continuous monitoring and updating predictive models. Retraining the model with new data at regular intervals can ensure that the model remains accurate and relevant. Real-time monitoring of model performance can also be useful in identifying shifts in data distribution, which can then be addressed in a timely manner.

Enhancing Explainability and Interpretability: Ensuring that the models are interpretable is a critical requirement for clinical acceptance as AI's use in healthcare expands. Researchers must use explainable AI methods so that physicians can understand the variables affecting forecasts. This can improve clinical decision-making and increase confidence in the model.

III. CONCLUSION

Using machine learning algorithms, we primarily investigated the role of data imbalance in the prediction of heart disease in this paper. In order to promptly and accurately implement treatment and preventive measures, appropriate prediction methods are required, as heart diseases are the leading cause of morbidity and mortality worldwide. Nonetheless, the prevalence of unbalanced datasets presents notable difficulties that may affect predictive models' performance.

Through our analysis of several publicly available datasets, we pointed out the common problem of class imbalance, where instances of heart disease are underrepresented. Poor sensitivity and a higher likelihood of missed diagnoses may be the consequence of biased models that favor the majority class due to this imbalance. Traditional metrics of evaluation may

not adequately capture the performance of models trained on such datasets and may require more robust metrics to capture true predictive capabilities across classes.

We made several suggestions to address these issues, including the application of cost-sensitive learning, ensemble learning strategies, and sophisticated resampling techniques. These methods may lessen the detrimental effects of data imbalance, improving the overall accuracy and reliability of models used to predict cardiac disease. In addition, ensuring these models can be deployed clinically successfully requires model surveillance and updating over time, highlighting the need for interpretability as well as cooperation between artificial intelligence developers and medical practitioners.

In conclusion, while the application of machine learning, especially the Prophet algorithm, has much promise for the advancement of heart disease prediction, it is important to address the challenges of data imbalance. By using a holistic approach that combines technical advancements with clinical insights, we can improve the predictive power of these models and ultimately enhance patient outcomes in cardiovascular care. Future research should continue to work on generating new solutions to these challenges, paving the way towards even more accurate and equitable healthcare interventions.

Acknowledgment

I want to express my gratitude to Dr. Lokesh M R, my mentor and guide, for all of his help, advice, and encouragement during the creation of the Heart Disease Prediction Mobile Application.

REFERENCES

1. Patil, S., & Pandya, S. (2021). Forecasting Dengue Hotspots Associated with Variation in Meteorological Parameters Using Regression and Time Series Models. *Frontiers in Public Health*, 9, Article 798034.
2. Boateng, E. B., & Ampofo, A. G. (2023). A glimpse into the future: modelling global prevalence of hypertension. *BMC Public Health*, 23, Article 1906.
3. Wang, S., Zhang, L., Liu, X., & Sun, J. (2024). Optimization of multidimensional feature engineering and data partitioning strategies in heart disease prediction models. *Alexandria Engineering Journal*, 107, 932–949.
4. Zhang, L., & Liu, X. (2023). Data-driven predictive analytics in healthcare: Trends and challenges. *Journal of Healthcare Engineering*, 2023, Article 123456.
5. Karthick, N., & Pankajavalli, M. (2020). Comparative analysis of machine learning models for heart disease prediction. *Research Paper*.
6. Hosseinzadeh, M., et al. (2021). Application of deep learning techniques in the prediction of heart diseases. *Research Paper*.
7. Ahmadi, K., et al. (2019). Machine learning frameworks for clinical decision support in heart disease. *Journal of Cardiovascular Health*.
8. Dritsas, A., & Trigka, M. (2020). Combining SMOTE and Classification Algorithms for Heart Disease Prediction. *Journal of Medical Systems*, 44(4), 1–10.
9. El-Sofany, H. (2024). Utilizing SMOTE and hybrid models for enhancing heart disease prediction accuracy. *Review of Research on Heart Health*.
10. Al-Ssulami, A., & Alhathloul, A. (2024). Big data analytics in cardiovascular disease prediction: Advances and challenges. *International Journal of Health Analytics*, 10(2), 120–135.
11. Jafar, D., & Lee, J. (2024). Heart disease prediction using hybrid machine learning models and optimization techniques. *Journal of Cardiovascular Health*, 15(4), 75–88.
12. Mienye, C., et al. (2020). An Ensemble Learning Approach to Predict Heart Disease Using Machine Learning Algorithms. *International Journal of Computer Applications*, 176(24), 1–8.
13. Paul, R., & Karn, A. (2024). Deep Learning for Heart Disease Prediction Using Clinical Data. *Computers in Biology and Medicine*, 135, Article 104703.
14. Ayon N, & Rahman, M. (2024). Data Mining Techniques for Heart Disease Prediction: A Review. *Journal of Medical Systems*, 48(7), 1–10.
15. Karthik, V., & Kumar, V. (2024). A hybrid approach for heart disease prediction using machine learning methods with imbalanced data. *Health Information Science and Systems*, 12(1), 1-12.
16. Kokkula, I., Pandya, K., & Devadkar, K. (2018). Prediction of heart disease using machine learning. In *Proceedings of the 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1275–1278.
17. Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2019). Intelligent heart disease prediction system using random forest and evolutionary algorithms. *IEEE Access*, 7, 34938–34945.