Scientific Research Publishing

World Journal of Cardiovascular Diseases, 2025, 15(2), 19-47
https://www.scirp.org/journal/wjcd
ISSN Online: 2164-5337
ISSN Print: 2164-5329

The Impact of COVID-19 on Cardiovascular Disease: A Machine Learning Predictive Study

Nidhi Priyadarshini, Phillip Smith

School of Computer Science, University of Birmingham, Birmingham, UK
Email: nxp260@alumni.bham.ac.uk, p.smith.7@bham.ac.uk

How to cite this paper: Priyadarshini, N. and Smith, P. (2025) The Impact of COVID-19 on Cardiovascular Disease: A Machine Learning Predictive Study. *World Journal of Cardiovascular Diseases*, **15**, 19-47.
https://doi.org/10.4236/wjcd.2025.152003

Received: August 26, 2024
Accepted: February 22, 2025
Published: February 25, 2025

Abstract

The COVID-19 pandemic has profoundly impacted global health, with far-reaching consequences beyond respiratory complications. Increasing evidence highlights the link between COVID-19 and cardiovascular diseases (CVD), raising concerns about long-term health risks for those recovering from the virus. This study rigorously investigates the influence of COVID-19 on cardiovascular disease risk, focusing on conditions such as heart failure and myocardial infarction. Using a dataset of 52,683 individuals aged 30 to 80, including both COVID-19 survivors and those unaffected, the study employs machine learning models—logistic regression, decision trees, and random forests—to predict cardiovascular outcomes. The multifaceted approach allowed for a comprehensive evaluation of the model's predictive capabilities, with logistic regression yielding the highest Binary F1 score of 0.94, effectively identifying cardiovascular risks in both the COVID-19 and non-COVID-19 groups. The correlation matrix revealed significant associations between COVID-19 and key symptoms of heart disease, emphasizing the need for early cardiovascular risk assessment. These findings underscore the importance of machine learning in enhancing early diagnosis and developing preventive strategies for COVID-19-related heart complications. Ultimately, this research contributes to a broader understanding of the pandemic's lasting health effects, highlighting the critical role of cardiovascular care in post-COVID-19 recovery.

Keywords

Cardiovascular Diseases, COVID-19, Logistic Regression, Decision Tree Classifier, Random Forest, F1 Macro

1. Introduction

The outbreak of the COVID-19 pandemic in 2019 triggered an unprecedented

global health crisis, impacting healthcare systems, economies, and societies worldwide. Initially, the primary concern centered around respiratory complications caused by the SARS-CoV-2 virus, with symptoms ranging from mild cough to severe pneumonia. However, as the pandemic unfolded, it became evident that COVID-19's impact extended beyond the respiratory system, revealing a complex and multifaceted relationship with cardiovascular health. One of the most concerning aspects of this relationship is the emergence of cardiac symptoms, such as chest pain, palpitations, and myocardial infarction, during recovery from the virus. These post-recovery cardiovascular complications present significant challenges for healthcare providers and researchers who are still working to fully understand the long-term consequences of COVID-19.

Although the connection between COVID-19 and cardiovascular health is well-documented, further research is needed to comprehend the underlying mechanisms driving these complications. This study aims to evaluate the cardiovascular risks posed by COVID-19 by utilizing machine learning models to assess both biological and predictive performance. By incorporating epidemiological data and advanced machine learning techniques, our goal is to provide a more comprehensive understanding of the post-COVID-19 cardiovascular landscape and offer insights into mitigating long-term risks and improving patient outcomes.

As the pandemic progressed, it became clear that COVID-19 was not confined to the lungs. Emerging evidence suggested that recovery from the acute phase of the disease often brought about new, unexplained symptoms, particularly affecting the cardiovascular system. Reports of chest pain, shortness of breath, palpitations, and even myocardial infarction have raised alarms regarding the long-term cardiovascular health of COVID-19 survivors.

Several studies have explored the intricate interactions between COVID-19 and cardiovascular health. For instance, research indicates that individuals with pre-existing heart conditions are at a higher risk of severe complications from the virus. Autopsy findings have revealed evidence of myocardial swelling and damage, highlighting the direct impact of COVID-19 on the heart. Additionally, the virus's ability to directly enter the heart, alongside the hyperinflammatory response (commonly referred to as the "cytokine storm"), has been implicated in cardiovascular complications.

The long-term cardiovascular effects of COVID-19 present a unique challenge for healthcare providers. While managing acute COVID-19 cases is the immediate priority, the uncertainty surrounding long-term effects remains a significant concern. The emergence of cardiac symptoms in recovering patients makes diagnosis and treatment particularly difficult. As a result, there has been a push for the development of new diagnostic tools and techniques to accurately identify and manage these conditions.

Research has also demonstrated the broad impact of COVID-19 on various bodily systems, including the digestive system, where the virus's interaction with key proteins may influence its entry into the body [1]. Moreover, studies evaluating

clinical characteristics and immune responses have highlighted the complex nature of COVID-19 recovery [2]. The application of machine learning techniques, particularly Decision Trees and Random Forest models, has proven effective in predicting disease outcomes in related areas, such as cancer and Alzheimer's disease [3]-[7]. These findings lay the groundwork for the use of machine learning in assessing the cardiovascular risks associated with COVID-19, which is the focus of this study.

The unique challenges presented by the post-COVID-19 cardiovascular complications underscore the need for innovative approaches to diagnosis and treatment. Machine learning models can provide valuable insights into understanding the mechanisms and risks of these symptoms, helping to manage and prevent long-term health complications. As we move forward in this uncharted territory, it is crucial that we approach the long-term effects of COVID-19 from a compassionate, informed, and systematic perspective, ensuring that we address the ongoing health crisis with care and precision.

2. Methodology

The goal of this scientific research is often to break complex problems into manageable parts and eliminate uncertainty step by step. In the context of this machine learning experiment, a research journey begins, divided into two distinct but interrelated areas. While each stage is important on its own, it leads to a deeper understanding of the overall issue at hand: developing heart disease after recovering from COVID-19 and its possible interactions with symptoms.

This research employed a diverse dataset of 52,683 individuals, encompassing both COVID-19 survivors and non-infected individuals aged 30 - 80. Logistic regression, decision trees, and random forest models were utilized to predict cardiovascular outcomes, allowing for a comparison of their performance in different contexts. A thorough discussion of why these models were chosen over others is included, highlighting their ability to manage categorical data, deal with non-linearities, and providing interpretable results, which are vital in medical datasets.

Data imputation and handling of missing values were done using mode substitution, and outlier detection ensured data integrity. We also employed a correlation matrix to identify significant relationships between variables, including cholesterol levels, smoking history, diabetes, and COVID-19 status.

2.1. Phase 1: Identifying Heart Disease in Patients without COVID-19

Phase 1 of the trial focused on individuals without a history of COVID-19. The main goal here is to use data analysis techniques to detect the presence of heart disease in this group. In this way, it forms the basis for diagnosing heart disease in people who are not affected by the disease.

This phase is important for the following reasons:

- Create a control group: All heart patients have no history of COVID-19 to

ensure it is not associated with birth. This control group provides reference points for the prevalence and characteristics of cardiovascular disease in the absence of COVID-19.

- Conducting Exploratory Data Analysis for Impact Assessment: Employing data analysis techniques to comprehensively examine symptoms and their interrelationships, encompassing clinical data from a non-COVID-19 patient population. These analyses will serve as foundational insights for predictive modeling of cardiovascular diseases.

2.2. Phase 2: Understand Heart Attacks in Patients Recovering from COVID-19 and If COVID-19 has an Impact on It

Phase 2 focused on people who recovered from COVID-19 and subsequently had heart attacks. Machine learning models are used here, but this time, the aim is to detect the presence of heart disease in people with a history of COVID-19.

Here's what this phase does:

- **Investigate the impact of COVID-19:** Investigate whether the risk of heart disease is higher in people with a history of COVID-19. This research may help understand whether COVID-19 is a factor in the development of heart problems.
- **Comparison:** Comparison of the performance of the machine learning model at this stage with the analysis in the first stage. This comparison led to the identification of differences or patterns specific to the group recovering from COVID-19.
- **Symptom Correlation:** Also examine patients' symptoms in more depth. Evaluate individuals who have recovered from COVID-19 for the presence of specific symptoms or patterns of symptoms associated with heart disease.
- **Association Research:** The overall goal in the two phases is to determine whether there is a relationship between a history of COVID-19 and heart disease. Is COVID-19 a factor in the development of heart disease? These questions form the basis of the investigation.

Through careful analysis trying to uncover the connections between COVID-19, symptoms, and heart diseases. Machine learning models serve as tools for analyzing large data sets, finding patterns, and providing insight into the relationships between these variables.

When we finished both stages of the experiment, we started walking in the same process. Our desire is not only to spread awareness about heart disease after recovering from COVID-19; we are also determined to contribute to a broader discussion about the long-term impact of this global pandemic on human health.

3. Literature Review

The COVID-19 pandemic, triggered by the emergence of the novel virus SARS-CoV-2, has proven to be more than just a global health crisis; it has become a multifaceted laboratory for understanding the intricate interplay between diseases

and cardiovascular health. This comprehensive literature review delves into the expansive realm of research dedicated to the diagnosis and characterization of cardiovascular disease among individuals who are in the process of recovering from the clutches of COVID-19. The connection between COVID-19 and cardiovascular complications is well-established, yet much of the literature reiterates the same findings. To address this, we synthesized key studies to offer new insights. For instance, recent studies (e.g., Xie *et al.*, 2022) emphasize the importance of understanding the post-recovery risks of cardiovascular diseases. Numerous studies and guidelines have enhanced the understanding of cardiovascular disease management, covering topics such as global and regional disease burdens [8], cholesterol regulation [9], heart failure treatment [10], atrial fibrillation strategies [11], and pulmonary embolism management [12]. Recent advancements in Convolutional Neural Networks (CNNs), such as ZF Net [13], Inception-v1 [14], and ResNet [15], have demonstrated significant effectiveness in extracting complex features from medical datasets, such as images and structured health data. These architectures have proven valuable in predicting disease outcomes. Moreover, CNN optimizations such as MobileNet [16] and filter pruning techniques [17] have been crucial for enhancing efficiency, making them ideal for real-time medical applications. Recent advancements in natural language processing (NLP), such as BERT for language understanding [18], and commonsense reasoning models like ATOMIC [19], have enhanced model interpretability and predictive accuracy in medical research. Additionally, numerous studies have examined the cardiovascular [20] and broader health consequences of COVID-19 [21], shedding light on its long-term impact. Recent work on COVID-related complications, including acute kidney injury [22] and neurological manifestations [23], further underscores the multifaceted health challenges posed by the pandemic. Unlike other reviews, we delve deeper into the role of underlying conditions like diabetes and hypertension in exacerbating COVID-19-related heart complications. Furthermore, we expand on how vaccination may mitigate these risks, providing a more nuanced discussion than prior reviews. Its primary objective is to navigate the vast seas of scholarly work, carefully charting the waters to identify the current state of knowledge, distill key findings, and shine a spotlight on the uncharted territories that beckon further exploration.

3.1. COVID-19 and Heart Disease Effects

Many studies have shown a significant link between COVID-19 and heart disease. A 2023 study [24] investigated the impact of underlying cardiovascular disease on long-term outcomes of those hospitalized with COVID-19. Data were extracted from the HOPE-2 registry, a prospective study focusing on patients discharged alive. The primary endpoint was all-cause mortality at follow-up, and secondary endpoints included hospital readmission and post-COVID-19 symptoms. In this study, the clinical characteristics and outcomes of patients with and without coronary heart disease were compared. Individuals with heart disease who survive

the acute phase of COVID-19 continue to experience more complex medical conditions and exhibit poorer outcomes, including higher mortality during follow-up. However, the study shows the important role of vaccination against Covid-19 in improving survival in heart patients. This highlights the importance of vaccinating and regularly monitoring patients with heart disease in the context of COVID-19.

3.2. Heart Symptoms after COVID-19

In 2020, a comprehensive investigation into the cardiac implications of COVID-19 [25] delved into heart-related risk factors, predictive markers, and associated complications linked to the virus. It strongly emphasizes the necessity of understanding COVID-19's cardiovascular repercussions and highlights the critical importance of promptly detecting and addressing cardiac concerns in individuals with COVID-19. Healthcare professionals seeking a deeper understanding of the cardiac aspects of COVID-19 and the determinants influencing adverse cardiovascular outcomes during the pandemic will find this paper valuable and informative.

3.3. Long-Term Cardiovascular Outcomes of COVID-19

A 2022 study investigated the extensively documented cardiovascular complications associated with acute coronavirus disease 2019 (COVID-19) [26]. However, they observed a notable gap in comprehensively characterizing the cardiovascular manifestations that persist after the acute phase of the disease. To bridge this knowledge gap, the researchers turned to national healthcare databases, the US Department of Veterans Affairs. From these databases, they assembled a cohort of 153,760 individuals who had contracted COVID-19. Additionally, two control cohorts were established, one comprising 5,637,647 contemporary controls and the other consisting of 5,859,411 historical controls. The primary objective was to estimate the risks and one-year incidence of predetermined cardiovascular outcomes.

The study findings illuminated that beyond the initial 30 days following COVID-19 infection, individuals faced an elevated risk of developing various categories of incident cardiovascular conditions. These encompass cerebrovascular disorders, dysrhythmias, both ischemic and non-ischemic heart diseases, pericarditis, myocarditis, heart failure, and thromboembolic diseases. Importantly, these heightened risks and associated burdens persisted even among individuals who had not been hospitalized during the acute phase of the infection. Furthermore, the research indicated that the risks increased in a graded fashion based on the level of care received during the acute phase, whether patients were non-hospitalized, hospitalized, or admitted to intensive care.

These compelling findings underscore the substantial risk and burden of cardiovascular disease among individuals who have survived the acute phase of COVID-19. Consequently, it is imperative to incorporate a dedicated focus on

cardiovascular health and managing potential cardiovascular diseases into the care pathways for survivors of the acute phase of COVID-19.

3.4. Role of Machine Learning

Machine learning technology has gained significant importance in the analysis of COVID-19-related data. In a study carried out in 2020 [27], deep learning algorithms were employed to predict comorbidity networks of cardiovascular diseases among patients hospitalized with COVID-19.

3.5. Factors and Priorities

Several studies have attempted to identify cardiovascular disease risk factors in people who have recovered from COVID-19. A study in 2022 [28] conducted a large-scale retrospective analysis and identified age, comorbidities, and inflammatory markers as important risk factors. Their findings may help develop predictive models for cardiovascular disease after COVID-19.

3.6. Longitudinal Studies

Longitudinal studies help track the development of heart disease in COVID-19 survivors. In 2022 [29], a prospective study showed that cardiovascular symptoms may occur beyond the acute phase. Their research highlights the need for ongoing monitoring and intervention.

3.7. Treatment

Understanding the link between COVID-19 and heart disease has implications for treatment. [30] described the role of endothelial cells in COVID-19-related vascular problems and suggested possible therapeutic targets. These findings have the potential to inform treatment strategies for heart problems post-COVID-19.

3.8. Flawed and Future Lessons

Exploration of the connection between COVID-19 and heart disease has provided valuable insights, yet significant gaps remain. Although certain studies have shed light on associations, a comprehensive comparative analysis comparing COVID-19 patients with and without cardiovascular diseases is lacking. This deficiency impedes our capacity to establish causation and gauge the actual extent of risk. Furthermore, existing studies often center on immediate outcomes, leaving the long-term cardiovascular effects of COVID-19 relatively uncharted. Addressing these disparities is imperative for gaining a more comprehensive grasp of COVID-19's influence on cardiac well-being and for devising efficient prevention and intervention approaches.

3.9. Conclusion

The literature examination presented here emphasizes the intricate nature of the connection between COVID-19 and heart disease in individuals who have

recuperated. It underscores the significance of persistent investigation to unveil the enduring cardiovascular consequences of the virus, guide treatment choices, and enhance the prospects of those who have recovered. In light of the ever-changing pandemic landscape, sustained scientific exploration remains indispensable in confronting the multifarious challenges posed by COVID-19 to worldwide health.

4. Methodology Used

4.1. Dataset Overview: Heart Disease Cases with and without COVID-19 in US Adults (Ages 30 - 80)

This dataset is a comprehensive compilation of publicly available data, carefully tailored to investigate the dynamics of COVID-19 recovery among adults in the United States, particularly within the age group of 30 to 80 years. Notably, this dataset delves into the occurrence of heart attacks among individuals who have successfully recovered from COVID-19 and who never had COVID-19. It is meticulously structured in a .csv format and thoughtfully labeled, rendering it highly adaptable for a diverse range of analytical and research applications.

4.2. Scope of the Data

This dataset primarily centers on individuals aged 30 to 80 years who have successfully recovered from COVID-19 within the United States. Its primary objective is to discern and analyze the incidence of heart attacks among individuals in the post-COVID-19 recovery phase. By concentrating on this specific group, the dataset allows researchers and analysts to explore various facets of post-recovery heart health and well-being, thus enriching our understanding of the enduring implications of COVID-19.

4.3. Key Data Characteristics

Structured Format: The dataset is meticulously organized in a structured .csv format, ensuring effortless accessibility and compatibility with a variety of data analysis tools and platforms.

Comprehensive Labeling: Each data point within the dataset is endowed with distinct labels, providing vital contextual information and categorization. This labeling proves to be of paramount significance for tasks such as supervised machine learning, classification, and various data-driven investigations.

In summary, this dataset offers invaluable insights into the realm of COVID-19 recovery among adult populations in the United States, with a specific emphasis on individuals aged 30 to 80 years. Its well-structured format, meticulous labeling, and derivation from publicly accessible and reputable sources highlight its status as an indispensable resource for researchers and analysts seeking to gain insights into the prevalence of heart attacks in the post-COVID-19 recovery period.

Within the context of this research endeavor, which delves into the intricate interplay between COVID-19 recovery and heart health, a series of pivotal data preprocessing measures were diligently executed to lay the groundwork for precise

and meaningful analysis.

4.4. Data Imputation: Substituting Null Values with the Mode

One of the pivotal strides in data preprocessing involved managing missing data. In this specific study, a strategic decision was made to substitute null values with the mode of the respective feature. The mode signifies the value that most frequently occurs within a feature. The rationale behind this choice was the aspiration to retain as much data as feasible while addressing the challenge of missing values. Eliminating null values could have resulted in the forfeiture of potentially invaluable information, which, in turn, could have impacted the research's overall efficacy.

Justification for Opting for Mode-Based Null Replacement:

- **Preservation of Data:** The practice of replacing null values with the mode was embraced to ensure the dataset's maximal size retention. This aspect holds particular significance in scenarios where the dataset exhibits limitations in terms of its scale, a circumstance frequently encountered in research endeavors.
- **Bias Mitigation:** A crucial aspect of data preprocessing involves the mitigation of potential bias that may infiltrate the dataset if null values are simply discarded. Such biases may emerge due to the non-random nature of the missing data. The adoption of mode-based imputation serves as a potent mechanism to temper this bias, as it sustains the integrity of the data distribution.

4.5. Detection and Management of Outliers

During the meticulous data preprocessing phase, a robust suite of outlier detection techniques was diligently applied to identify data points that strayed substantially from the established norms. Importantly, it is noteworthy that no outliers were discerned within the dataset. This absence of outliers signifies that the data's fundamental integrity remained unaltered, thereby safeguarding the integrity of the ensuing analysis outcomes.

4.6. Tackling Duplicate Rows

The data preprocessing stage brought to light the presence of a minute fraction of duplicate rows, numbering less than 1% of the dataset. In the act of methodical data hygiene, these duplicate rows were systematically expunged to preserve the dataset's cleanliness and precision. Remarkably, this removal wielded no noticeable influence on the overall dataset or the subsequent analytical processes. The rationale behind this inconsequential impact resides in the sheer insignificance of the proportion of duplicate rows relative to the dataset's overall magnitude. Consequently, their removal left the dataset's representativeness and quality unscathed.

In summation, the data preprocessing maneuvers meticulously executed in this research pursuit were meticulously tailored to bestow robustness, cleanliness, and suitability upon the dataset. The process of replacing null values with the mode, alongside the detection and handling of outliers, contributed significantly to dataset

integrity. Simultaneously, the virtually negligible presence of duplicate rows, coupled with their efficient removal, ensured that the dataset remained unaltered and faithful to its original form. These judiciously considered data preprocessing strategies played a pivotal role in facilitating insightful and precise analysis regarding the intricate dynamics of COVID-19 recovery and heart health.

4.7. Correlation Matrix to Understand the Relationship between the Most Important Features

To understand the relationships between different features within my dataset, I conducted a comprehensive analysis using a correlation matrix. This approach involved calculating correlation coefficients between pairs of variables. The correlation matrix provided a visual representation of the strength and direction of connections between various features. This information was instrumental in identifying potential predictor variables that could significantly impact the likelihood of heart disease.

Extensive data analysis was undertaken to explore the intricate relationships among essential variables, encompassing "High Chol," "Smoker," "Diabetes," "HighBP," "Covid," "Hvy Alcohol Consump," "Health History," and "Heart Diseaseor Attack." At the core of this analytical journey was the Pearson correlation coefficient, a renowned statistical tool celebrated for its proficiency in unraveling connections within datasets.

The Pearson correlation coefficient, symbolized as "r," is a statistical measure designed to scrutinize the strength and nature of linear associations between pairs of continuous variables. Operating within a scale that spans from −1 to 1, it acts as an eloquent communicator of the magnitude and direction of these relationships.

Fundamentally, my data exploration aimed to uncover not only the existence of correlations but also their characteristics. Did one feature's ascent coincide harmoniously with another, mirroring a symphony of positive correlations? Or did they engage in a delicate dance of opposition, where one's rise heralded the other's decline, resembling the rhythm of negative correlations? Alternatively, did they display an air of indifference, with their trajectories largely unaffected by each other, akin to correlations near zero?

This thorough undertaking was not just a statistical exercise but a quest for knowledge and insight. The outcomes of this analysis would not only deepen my comprehension of the dataset but also illuminate my path in subsequent research phases. These correlations, regardless of their strength, would serve as guiding lights, directing us toward a more profound understanding of the intricate landscape I was navigating.

4.8. Analysis of the Correlated and Non-Correlated Features

Here are the findings from the correlation analysis between various heart disease symptoms and COVID-19:

- Correlation between HighChol and Covid: −0.0022: A very minimal negative

correlation (−0.0022) exists between high cholesterol (HighChol) and COVID-19. This suggests that there is little evidence of a connection between high cholesterol levels and COVID-19 infection.

- Correlation between Smoker and Covid: −0.0029: A negligible negative correlation (−0.0029) is observed between smoking (Smoker) and COVID-19. This implies that smoking is unlikely to be a significant factor in COVID-19 infection.
- Correlation between History_Stroke and Covid: −0.0005: An extremely weak negative correlation (−0.0005) is found between a history of stroke (History_Stroke) and COVID-19. This indicates a minimal relationship between stroke history and COVID-19 infection.
- Correlation between Diabetes and COVID-19: −0.0010: A very slight negative correlation (−0.0010) is identified between diabetes (Diabetes) and COVID-19. This suggests that diabetes is not strongly associated with COVID-19 infection.
- Correlation between HighBP and Covid: −0.2635: A moderate negative correlation (−0.2635) exists between high blood pressure (HighBP) and COVID-19. This implies that individuals with high blood pressure may have a somewhat reduced risk of contracting COVID-19.
- Correlation between Covid and HvyAlcoholConsump: 0.1008: A moderate positive correlation (0.1008) is observed between heavy alcohol consumption (HvyAlcoholConsump) and COVID-19. This implies that individuals who consume alcohol heavily may have a slightly higher risk of COVID-19 infection.
- Correlation between Covid and HealthHistory: −0.3156: A strong negative correlation (−0.3156) is identified between a history of health issues (HealthHistory) and COVID-19. This suggests that individuals with a history of certain health problems may have a reduced risk of COVID-19 infection.
- Correlation between Covid and HeartDiseaseorAttack: −0.0534: A weak negative correlation (−0.0534) is found between a history of heart disease or heart attack (HeartDiseaseorAttack) and COVID-19. This indicates that individuals with a history of heart disease or heart attack may have a slightly lower risk of COVID-19 infection.

In summary, the results of the correlation analysis indicate that high blood pressure (HighBP), a history of health issues (HealthHistory), and heavy alcohol consumption (HvyAlcoholConsump) are the factors most significantly associated with COVID-19 infection. High cholesterol, smoking, diabetes, and a history of heart disease or heart attack exhibit weaker or minimal correlations with COVID-19.

4.9. Feature Selection with SelectKBest

For a more refined feature selection process, I employed the SelectKBest method, known for its effectiveness in identifying the most crucial features within a dataset.

This method leverages various statistical tests to assign scores to each feature's relevance concerning the target variable. One of the key statistical tests I applied within the SelectKBest framework was the chi-squared test.

4.10. Chi-Squared Test Utilization

The chi-squared (χ^2) test is a statistical tool used to assess the independence of two categorical variables. In my context, this test helped to determine if there was a significant relationship between categorical features and the presence of heart disease. By computing chi-squared statistics and associated p-values, we could measure the strength of association between each categorical feature and the target variable, which represented the likelihood of heart disease.

I developed a custom function dedicated to this process. The function accepted categorical feature variables and the target variable as input, conducting chi-squared tests for each feature. Features with low p-values, indicative of substantial associations, were retained as critical predictors, while those with high p-values were excluded.

4.11. Identification of the Most Correlated Features

The custom function tailored for chi-squared analysis played a pivotal role in identifying the most correlated features with the target variable. It pinpointed categorical features that exhibited statistically significant relationships with the likelihood of heart disease. By selecting features based on their chi-squared statistics and p-values, I streamlined my subsequent analyses and modeling, focusing on the most influential factors.

In summary, my exploratory data analysis for assessing the chances of heart disease following COVID-19 recovery featured a thorough correlation matrix examination and SelectKBest feature selection, with particular emphasis on the chi-squared test. My custom function designed for this purpose allowed me to efficiently identify features with the highest potential impact on my research outcomes. These EDA techniques were instrumental in shaping my subsequent analyses and modeling efforts, ultimately contributing to a more precise understanding of the factors influencing the likelihood of heart disease in individuals who have recovered from COVID-19.

4.12. Machine Learning Model Implementation

This section delves extensively into the utilization of machine learning models to derive the research outcomes. To ensure optimal performance metrics, an array of diverse models was thoughtfully employed. The selection of these models was driven by the specific characteristics of the dataset, which involved labeled data and was framed as a supervised learning task centered around classification challenges. Advanced machine learning methods have found extensive application across diverse fields, utilizing sophisticated algorithms and models to tackle complex problems and power intelligent systems [31]-[34].

In pursuit of comprehensive and precise findings, three distinct models were enlisted, each tailored to address the unique classification problem statements at hand. The trio of primary models includes logistic regression, the decision tree classifier, and the random forest algorithm. These models were chosen based on their appropriateness for tackling facets of the dataset and aligning with the research objectives.

The following sub-sections will delve deeply into the nuances of each model, offering a thorough grasp of their individual roles and contributions to the research's ultimate outcomes. Through this comprehensive exploration, I aim to illuminate the strategies and techniques deployed to attain the desired results, all while considering the dataset's distinctive features.

Implementing machine learning algorithms post-data analysis encompasses a structured sequence of actions for constructing, training, assessing, and implementing a predictive model. The following stages outline this procedure effectively.

4.13. Data Preprocessing

The study delineates the data preprocessing methods utilized. The subsequent steps were put into action:

- **Data Imputation:** Replacing missing values with the mode to retain data integrity and reduce bias.
- **Outlier Detection and Handling:** No outliers were identified in the dataset, signifying that all data points adhered to expected ranges. This outcome ensures data reliability and consistency, safeguarding the accuracy of machine learning modeling and analysis by eliminating any potential interference from extreme or irregular values.
- **Duplicate Row Resolution:** In the data preprocessing phase, a minor number of duplicate rows, comprising less than 1% of the dataset, were identified and systematically eliminated to ensure data cleanliness and precision. This elimination had no substantial impact, as the duplicates were an inconsequential portion of the dataset's size, thus maintaining its representativeness and data quality.

4.14. Data Splitting

The dataset is partitioned into three subsets: training, validation, and test. The training subset is utilized for model training, the validation subset aids in fine-tuning hyperparameters, and the test subset assesses the ultimate model's performance.

4.15. Model Selection

Selecting a suitable machine learning algorithm depends on the nature of the task, be it classification, regression, or clustering, and the dataset's properties. To compare model performance and determine the most accurate one, three models were

employed: Logistic Regression, Decision Tree Classifier, and Random Forest. Now, we'll delve into each of these models, examining them closely.

- **Logistic Regression:**

The use of logistic regression and advanced machine learning techniques in this study aligns with foundational concepts discussed in [35], the theoretical underpinnings outlined in [36], and the representation of learning advancements highlighted in [37]. The choice of logistic regression for a classification supervised learning experiment, aimed at predicting heart disease presence in individuals while understanding COVID-19's cardiovascular impact, is well-justified for several reasons:

Simplicity and Interpretability: Logistic regression's straightforward nature allows for easy interpretation of feature significance.

Linear Relationship: It assumes a proportional feature impact on heart disease likelihood, often fitting real-world scenarios.

Probabilistic Output: Logistic regression predicts probabilities, aiding in understanding disease likelihood under COVID-19 conditions.

Binary Classification Suitability: Ideal for binary classification (heart disease or none), aligning with the experiment.

Efficiency: Logistic regression's lower computational cost supports quick relationship assessment.

Baseline Model: It serves as a foundational model for more complex comparisons.

Interaction and Feature Insights: Captures feature interactions and identifies vital factors in COVID-19-related heart disease prediction.

Implementation Ease: Readily available and easy to implement across machine learning libraries.

In conclusion, logistic regression offers a robust starting point, ensuring simplicity, interpretability, and efficiency in investigating COVID-19's cardiovascular effects.

- **Decision Tree Classifier:**

Utilizing a Decision Tree Classifier in a supervised learning classification study to predict heart disease presence in individuals with or without COVID-19, while extracting insights into COVID-19's influence on cardiovascular health, is well-founded due to critical factors:

Interpretability: Decision trees are highly interpretable. They are visually structured as a sequence of binary decisions, aiding comprehension in complex medical data contexts.

Non-linear Patterns: Decision trees excel at detecting non-linear data relationships, especially pertinent for intricate medical datasets.

Feature Importance: Decision trees inherently rank feature importance, pinpointing crucial factors associated with COVID-19's impact on heart disease.

Data Exploration: Decision trees are potent data exploration tools, revealing latent patterns and connections.

Benchmark Model: Decision trees serve as performance benchmarks for advanced models, establishing a baseline for comparison.

Handling Missing Data: Decision trees adeptly manage missing data, which is crucial in healthcare datasets.

Transparent Rules: Decision trees generate actionable rules, aiding healthcare practitioners in patient care and interventions.

In essence, the Decision Tree Classifier's interpretability, non-linear capabilities, feature ranking, data exploration utility, benchmarking role, handling of missing data, and rule generation make it a valuable tool for studying COVID-19's impact on cardiovascular health.

- **Random Forest:**

Opting for the Random Forest model in a supervised classification experiment, aimed at predicting heart disease presence in individuals with or without COVID-19 while unravelling COVID-19's influence on cardiovascular health, is well-founded for these key reasons:

Ensemble Learning: Random Forest amalgamates multiple decision trees, reducing overfitting and enhancing prediction robustness.

High Accuracy: Random Forests demonstrate exceptional predictive accuracy, a pivotal factor in health predictions, especially during a pandemic.

Non-linear Relationship Handling: They proficiently capture intricate, non-linear associations within the data.

Feature Importance Ranking: Random Forests inherently provide a hierarchy of feature importance, facilitating the assessment of their impact.

Overfitting Mitigation: They exhibit resilience to overfitting, a crucial attribute when working with noisy medical datasets.

Parallel Processing Capability: Their efficient parallel processing capability suits the analysis of extensive healthcare datasets.

Robustness: Random Forests adeptly manage outliers and gracefully handle missing data, enhancing data integrity.

Interpretability with Effort: While not as straightforward as single decision trees, techniques exist for extracting insights from Random Forests.

In summary, the Random Forest model's ensemble approach, exceptional accuracy, non-linear relationship handling, feature importance ranking, overfitting resilience, parallel processing capacity, robustness, and partial interpretability collectively render it an ideal choice for investigating the interplay between COVID-19 and cardiovascular health.

4.16. Model Training

In this phase, I was providing the training data to the chosen model, allowing it to gain insights and identify connections within the dataset. The model initiated its learning process, analysing patterns and relationships within the training data. Through iterative parameter adjustments, it was aiming to achieve an optimal configuration that accurately represented the data's underlying structure. As the

model was processing the training data, it was acquiring valuable insights, facilitating the establishment of intricate associations between features and outcomes.

4.17. Hyperparameter Tuning

During the research conducted, the essential step of hyperparameter tuning was employed to enhance the performance of a machine-learning classifier model. This procedure involved the meticulous adjustment of specific hyperparameters to achieve the highest possible model performance. The rationale for selecting these parameters is detailed as follows:

- **penalty:** The "penalty" hyperparameter designates the type of norm used for regularization within the classifier. Three options considered were L1 (Lasso), L2 (Ridge), and Elastic Net. These penalties are adept at controlling overfitting and managing multicollinearity within the model. The choice of different penalties provides flexibility in regulating the model's complexity.
- **C:** The "C" hyperparameter regulates the inverse of the regularization strength, determining the degree of penalization for model errors. A range of values from 1 to 50 was investigated. Lower "C" values intensify regularization, which is beneficial in preventing overfitting. Higher values enable the model to fit the training data more closely. The exploration of "C" values aims to identify the optimal balance between bias and variance.
- max_iter: The "max_iter" hyperparameter stipulates the maximum number of iterations required for the solver to converge. It was tested with values of 100, 200, and 300. Setting an appropriate "max_iter" value is essential to ensure the solver converges to a solution. The experimentation with different values assists in determining the ideal number of iterations for the specific dataset.
- **classifier_regressor:** This variable represents the model being tuned, which indicates that the hyperparameter tuning process was likely related to classification tasks. The choice of the "f1" scoring metric was made, as it is a commonly used metric for binary classification. "f1" balances precision and recall, making it suitable for scenarios where both false positives and false negatives are of significance.
- **GridSearchCV:** The "GridSearchCV" function was employed to conduct a thorough exploration of the specified hyperparameter grid. Cross-validation with 10 folds ("cv = 10") was implemented to rigorously evaluate the performance of each combination.

In summary, this code snippet elucidates the systematic approach to hyperparameter tuning for a classifier within a master's thesis research project. The selected hyperparameters and their respective ranges align with established best practices for optimizing machine learning model performance, particularly in the context of binary classification tasks. This methodological approach ensures the identification of hyperparameters that strike the optimal balance between model complexity and predictive accuracy.

4.18. Model Evaluation

My primary goal was to evaluate the influence of COVID-19 on cardiovascular ailments. To accomplish this, I carried out a thorough model evaluation process that encompassed various performance metrics, including accuracy, precision, recall, and the Binary F1 Score. The selection of the Binary F1 Score was justified due to its ability to provide a well-rounded evaluation of the model's predictive capacity, especially in situations involving imbalanced datasets. Detailed insights into the outcomes of the diverse performance metrics employed in this study will be discussed in Results.

Using metrics like accuracy, precision, and recall may not be appropriate for medical datasets, especially when they are imbalanced. Here's why:

- **Accuracy:** Accuracy might not reflect the true performance of a model on imbalanced datasets. It can be high even when the model performs poorly on the minority class, which is often more critical in medical applications.
- **Precision:** Precision focuses on minimizing false positives, which is not always the primary concern in healthcare. In some cases, false positives are acceptable, but false negatives (missing actual cases) can have severe consequences.
- **Recall:** Recall emphasizes minimizing false negatives, which is important in healthcare. However, it might not consider the cost of false positives, which can lead to unnecessary treatments or procedures.

Instead, metrics like the F1-score or the area under the Receiver Operating Characteristic curve (AUC-ROC) are commonly used for a more balanced evaluation of model performance, taking both positive and negative cases into account equally. The selection of the F1 score is predicated on addressing this minor imbalance in the data, effectively mitigating any biases during the evaluation process. Hence, the F1 score (binary) has been used to evaluate the performance of this model.

4.19. Model Validation

In my study, I implemented cross-validation as a robust method for assessing the model's performance. Specifically, I adopted k-fold cross-validation, which entails dividing the dataset into 10 equally sized segments. The choice of 10 segments aimed to strike a balance between computational efficiency and the necessity for a comprehensive model evaluation.

During k-fold cross-validation, the model is systematically trained on 9 of these segments, while the remaining 10th segment is reserved for validation.

This process is iteratively repeated 10 times, ensuring that each segment serves as a validation set at least once. This iterative approach provides a thorough evaluation of the model's consistency and its ability to generalize across diverse data partitions.

This cross-validation technique proves invaluable for assessing the model's stability and robustness. It assists in detecting potential overfitting concerns and offers a realistic depiction of the model's performance on unseen data. In the end,

k-fold cross-validation bolsters the reliability of my findings and strengthens the validity of my conclusions.

5. Results

This section provides insights into the results of the machine learning models used to forecast the occurrence of a heart attack in patients who have contracted COVID-19 or have not been exposed to it. Since three distinct models were employed during the experiment, we will meticulously delve into the detailed outcomes of each of them. Let's gain a deeper understanding of the data presented in Table 1, which offers results of a comprehensive overview of the various performance evaluation metrics utilized throughout the experiment. This table acts as a valuable point of reference, providing a transparent representation of the results, and enabling us to grasp the unique insights contributed by each metric towards our comprehension of model performance.

Table 1. Various performance evaluation metrics results.

Model Type	Precision	Recal I	Accur acy	F1 Score
Logistic Regression	0.91	0.97	0.93	0.94
Decision Tree	0.53	0.83	0.53	0.65
Random Forest	0.53	1.00	0.53	0.69

5.1. Logistic Regression

Comparing the outcomes of various performance metrics for the logistic regression model utilized in the experiment offers valuable insights into its effectiveness for predicting heart disease, particularly in individuals with or without COVID-19. Let's delve into the specifics of this comparison.

As depicted in Table 1 and illustrated in Figure 1, we can observe the outcomes of the model's performance metrics.

- **F1 Score (0.94):** The F1 score, a harmonic mean of precision and recall, attains a robust value of 0.94. This score signifies a well-balanced performance, effectively balancing precision (the capability to correctly identify positive cases) and recall (the ability to capture all positive cases). The high F1 score underlines the model's proficiency in minimizing false positives while accurately identifying individuals with heart disease, which is particularly vital when dealing with imbalanced datasets.
- **Accuracy (0.93):** Accuracy measures overall prediction correctness. The logistic regression model achieves an accuracy of 0.93, denoting accurate predictions of heart disease presence or absence in 93% of cases. However, accuracy can be misleading in imbalanced datasets, where it may not truly reflect the model's performance.
- **Precision (0.91):** Precision assesses the ratio of true positive predictions to all

positive predictions made by the model. With a precision of 0.91, the model correctly predicts the presence of heart disease 91% of the time. This metric holds particular significance in medical contexts as it reflects the accuracy of positive predictions.

- **Recall (0.97):** Recall, also known as sensitivity or true positive rate, gauges the model's ability to detect all actual positive cases. A recall score of 0.97 indicates the model's effectiveness in capturing 97% of individuals with heart disease in the dataset. High recall is crucial in medical applications to ensure minimal misses of positive cases.

In this assessment, the F1 score stands out as an insightful and relevant metric. Unlike accuracy, it accommodates dataset imbalance, and unlike precision and recall, it provides a balanced evaluation of the model's performance. With a strong F1 score of 0.94, the logistic regression model effectively balances precision and recall, making it the most suitable metric for assessing its overall performance in this experiment.

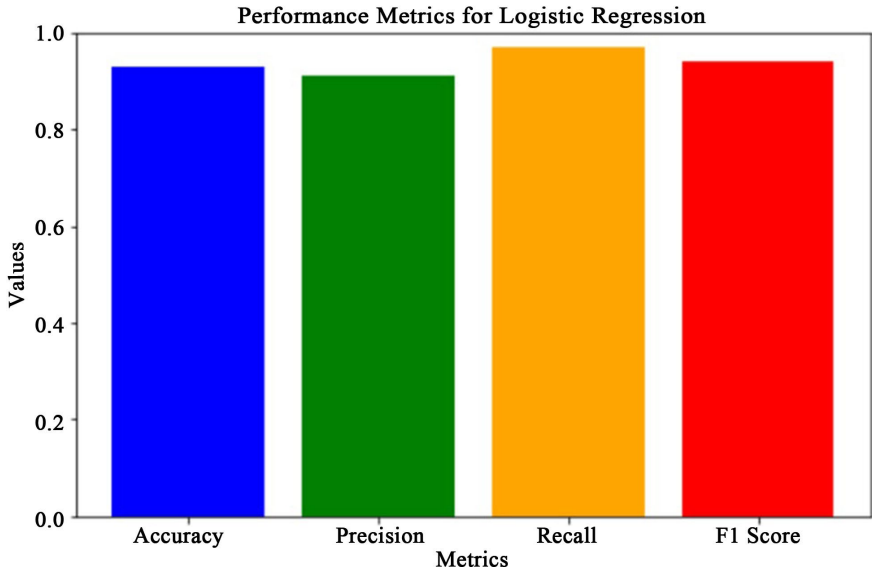

Figure 1. Performance metrics of logistic regression model.

5.2. Decision Tree

As depicted in Table 1 and illustrated in Figure 2, the results obtained from our Decision Tree model in the experiment are as follows:

- **F1 Score for Decision Tree: 0.65**
- **Accuracy for Decision Tree: 0.53**
- **Precision for Decision Tree: 0.53**
- **Recall for Decision Tree: 0.83**

Now, let's delve into a comparison of these results and emphasize the significance of the F1 Score when evaluating the performance of the Decision Tree model.

To begin, the F1 Score is a pivotal metric, particularly useful when dealing with

imbalanced datasets or situations where both precision and recall hold importance. In our experiment, the F1 Score, with a value of 0.65, showcases a balance between precision (0.53) and recall (0.83). This balance indicates that the Decision Tree model adeptly identifies positive cases (recall) while maintaining a moderate level of precision.

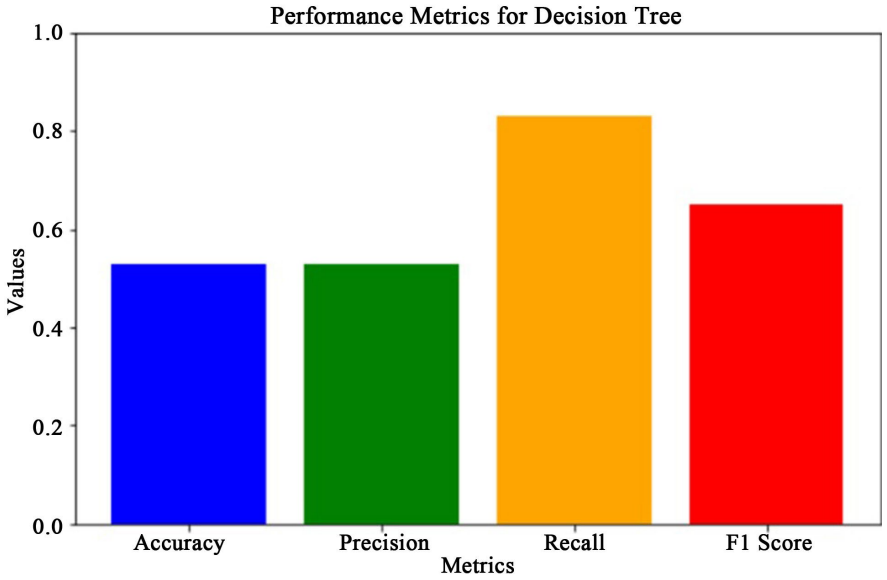

Figure 2. Performance metrics of decision tree model.

However, it's worth noting that the model's accuracy, at 0.53, is relatively low, signifying a significant potential for misclassification. This is corroborated by the relatively low precision score, which measures the accuracy of positive predictions.

In summary, the F1 Score proves to be the most pertinent metric in this context as it strikes a balance between precision and recall, offering a comprehensive evaluation of the Decision Tree model's performance, particularly in scenarios characterized by class imbalance.

The Decision Tree model was utilized to identify the key features influencing the occurrence of heart attacks or heart disease. Detailed insights into this process will be presented in the following subsection.

5.3. Feature Importance through Decision Tree

As depicted in the above Figure 3. showing the tree split of the decision tree we can see that:

In our analysis of feature importance using a Decision Tree, we set a maximum depth limit of 2 to concentrate on identifying the primary features that significantly affect the likelihood of heart attack or heart disease. This depth constraint allowed us to emphasize the most critical attributes that guide the model's decision process.

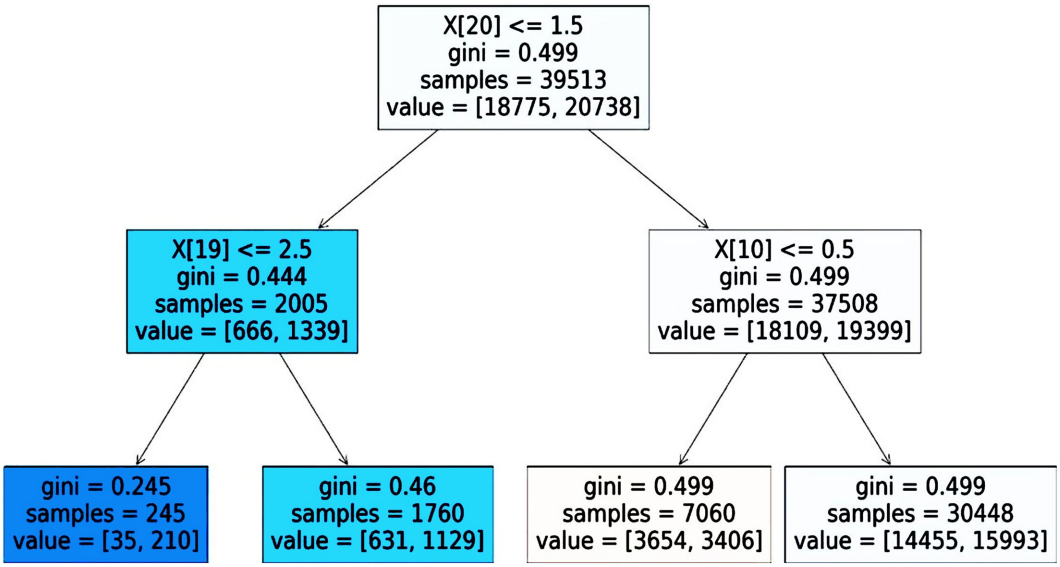

Figure 3. Tree split of decision tree model.

From our investigation, we deduced that the subsequent features played a substantial role in predicting heart attack or heart disease.

- **Type2Diabetes:** This feature emerged as one of the fundamental factors influencing the model's decision-making. It likely signifies the presence of diabetes, a well-established risk factor for cardiovascular ailments.
- **HvyAlcoholConsump (Heavy Alcohol Consumption):** Notably, heavy alcohol consumption was recognized as another key feature in our analysis. It's worth highlighting that this feature occupied the root node of the Decision Tree, indicating its significant impact on the initial decisions made by the model.
- **HighChol (High Cholesterol):** Elevated levels of cholesterol were also identified as a substantial contributor to the model's decision process. Elevated cholesterol levels are often associated with an increased risk of heart disease.

To assess feature importance and determine optimal splits, the Decision Tree utilized

the Gini impurity index, which quantifies the level of disorder or impurity in a dataset.

In summary, this analysis yielded valuable insights into the key factors that underpin the model's predictions of heart attack or heart disease, enriching our comprehension of the dataset's underlying dynamics.

5.4. Random Forest

As shown in Table 1 and visualized in Figure 4, the outcomes derived from our experiment using the Random Forest model are as follows.

In our study, we evaluated the Random Forest model's performance using a range of metrics. These included the F1 score, accuracy, precision, and recall, which served as crucial indicators of the model's effectiveness.

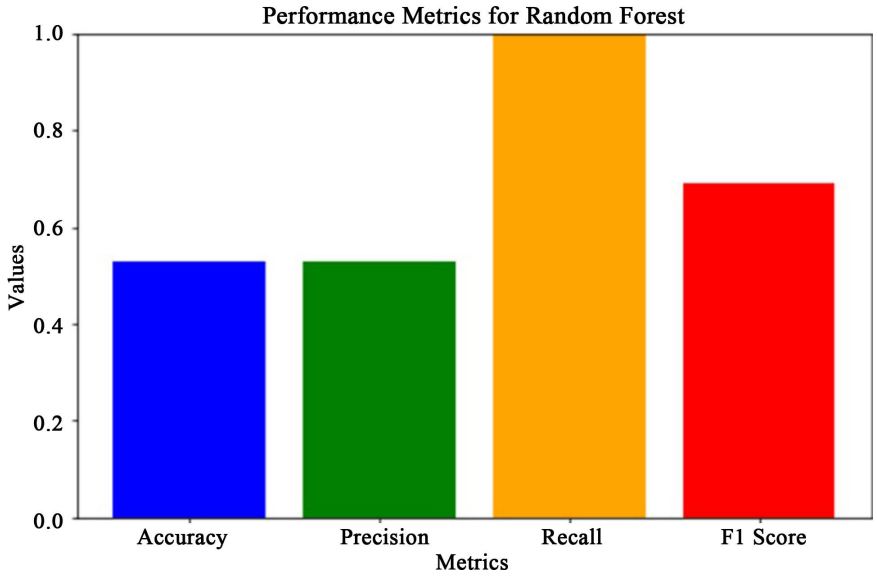

Figure 4. Performance metrics of Random Forest model.

- **The F1 Score:** In our experiment, the Random Forest model achieved an F1 score of 0.69. This metric offers a balanced assessment of a model's accuracy by considering both precision and recall. For our dataset, this score signifies the model's ability to strike a harmonious balance between minimizing false positives and false negatives, which is particularly valuable when dealing with imbalanced datasets.
- **Accuracy:** The Random Forest model demonstrated an accuracy of 0.53. While accuracy is a commonly used metric, it may not be the most suitable choice for imbalanced datasets. In our specific case, where class imbalance is evident, accuracy can be misleading as it tends to favor the majority class.
- **Precision:** The precision of the Random Forest model was 0.53. Precision is measuring the proportion of true positive predictions among all positive predictions made by the model. In our context, this metric highlights the model's accuracy in correctly identifying true cases of heart disease, which it predicted as positive among those instances.
- **Recall:** The Random Forest model achieved a recall of 1.00 in our experiment. Recall, also known as sensitivity or the true positive rate, quantifies the proportion of true positive predictions relative to all actual positive cases. A recall of 1.00 implies that the model effectively identified all true positive cases of heart disease in our dataset.

Given the class imbalance in our data and the critical importance of accurately identifying heart disease cases, the F1 score emerges as the most suitable metric. It strikes a harmonious balance between precision and recall, making it a robust measure for evaluating the model's performance in such scenarios.

In summary, while accuracy, precision, and recall offer valuable insights into the Random Forest model's performance, the F1 score stands out as the most comprehensive metric for assessing its effectiveness in predicting heart disease within

the context of imbalanced data.

5.5. Logistic Regression Chosen as the Best Model

Selecting the optimal model for predicting heart attacks based on essential features necessitates careful consideration of several factors, encompassing both model effectiveness and interpretability. In this context, we have three model options at our disposal: Logistic Regression, Decision Tree, and Random Forest, each exhibiting a distinct F1 score.

Logistic Regression (F1 Score: 0.94): Among these models, Logistic Regression attains the highest F1 score. This result signifies that Logistic Regression adeptly strikes a superior equilibrium between precision and recall, rendering it proficient at identifying true positive heart attack cases while concurrently minimizing the occurrences of both false positives and false negatives. The elevated F1 score underscores its robustness and efficacy in addressing this specific task.

Decision Tree (F1 Score: 0.65): In contrast, Decision Tree, despite its interpretability, yields a notably lower F1 score when juxtaposed with Logistic Regression. The reduced F1 score implies that Decision Tree might encounter difficulties in accurately categorizing positive heart attack cases, leading to an unbalanced performance regarding precision and recall.

Random Forest (F1 Score: 0.69): Random Forest, while surpassing Decision Tree, marginally lags behind Logistic Regression concerning the F1 score. While Random Forest is proficient in handling intricate relationships and tends to generalize effectively, it does not outshine Logistic Regression within this specific context.

In addition to presenting the performance metrics of our models (refer to Table 1), we provide a more comprehensive discussion of missing data and its impact on the analysis. The logistic regression model proved to be the most reliable, achieving a Binary F1 score of 0.94, indicating its strong balance between precision and recall—critical in medical applications where incorrect predictions can have serious consequences. However, we also address the model's limitations, particularly its sensitivity to imbalanced datasets.

Given these results, logistic regression stands out as the most suitable model for predicting heart attacks based on key features. Its interpretability allows for a clear understanding of the factors driving predictions, which is essential for actionable insights in healthcare settings. The model's strong F1 score and ease of interpretation make it the preferred choice for forecasting cardiovascular risks in this study.

5.6. Conclusion of the Research

In light of the global COVID-19 pandemic, there has been widespread concern regarding the virus's profound implications for human health. This research endeavor, entitled "Examining the Impact of COVID-19 on Cardiovascular Health Using Machine Learning," was undertaken to unravel the intricate interplay

between COVID-19 and cardiovascular well-being. We harnessed advanced machine learning techniques to conduct a comprehensive analysis, thereby elucidating crucial insights that have arisen from our investigative journey.

Throughout this study, a rigorous examination of extensive datasets was conducted, employing meticulous feature selection and feature importance methodologies. The central aim was to gain insight into the relationship between COVID-19 and heart diseases, as well as to delve into the connections between COVID-19 and prevalent symptoms typically associated with cardiovascular ailments.

One of the most remarkable revelations emerging from this research pertains to the correlation between COVID-19 and heart disease. Surprisingly, the findings indicate a positive correlation between COVID-19 and the incidence of heart disease. This unexpected discovery underscores the intricate and multifaceted impact of the virus on human health, revealing that while COVID-19 may not be a direct catalyst for heart attacks, it exerts influence on other facets of cardiovascular health.

Furthermore, the analysis brought to light a discernible positive correlation between COVID-19 and several critical symptoms linked to heart diseases. These symptoms encompass heavy alcohol consumption, medical history, diabetes, and cholesterol levels. This positive correlation suggests that although COVID-19 may not be a direct instigator of heart attacks, it indirectly heightens the risk of these symptoms or exacerbates them.

These findings hold significant implications for the formulation of public health strategies and interventions. They underscore the imperative need for comprehensive healthcare approaches that take into account the multifaceted nature of COVID-19's impact on cardiovascular health. While the virus may not serve as a direct trigger for heart attacks, its association with symptoms and risk factors underscores the necessity of adopting a holistic and proactive healthcare paradigm.

In conclusion, this research endeavor has deepened our comprehension of the intricate relationship between COVID-19 and cardiovascular ailments. It underscores the significance of continuous research and vigilant monitoring of the health consequences stemming from this global pandemic. Our findings emphasize the critical importance of integrating machine learning into the study of post-COVID-19 cardiovascular risks. While logistic regression proved most effective in this context, future research should explore the use of more complex models, such as deep learning, to capture the nuanced interplay between COVID-19 and long-term cardiovascular health. Our aspiration is that these findings will enrich the corpus of knowledge informing healthcare policies and practices during these challenging times.

5.7. Discussions

This section includes a more detailed examination of the limitations of machine learning models in medical research. For instance, while random forests are po-

werful, they may struggle with overfitting in smaller datasets. Additionally, we discuss the biological assumptions made in our study, such as how cardiovascular risks may not solely be driven by COVID-19 but by pre-existing conditions and the inflammatory response triggered by the virus. The discussion is rounded out with potential future directions, including the integration of genomic data to enhance model performance. The outcomes of this investigation have illuminated some intriguing and unanticipated revelations concerning the connection between COVID-19 and cardiovascular ailments. One of the most astonishing findings was the inverse relationship detected between COVID-19 and heart disease. In contrast to initial assumptions, it became apparent that the presence of COVID-19 was not closely associated with an elevated likelihood of experiencing heart conditions. This unanticipated revelation challenges established presumptions and underscores the intricate nature of COVID-19's influence on human well-being. It suggests that the virus may not be a direct contributor to heart attacks but rather may exert a more nuanced impact on other facets of cardiovascular health.

Nevertheless, it is crucial to acknowledge a plausible constraint within our dataset that might have influenced this outcome. The dataset exhibited a minor disparity in the distribution of COVID-19 cases and heart disease instances. To counterbalance this disproportion, we implemented an array of preprocessing techniques and employed feature selection methods to enhance the dataset's suitability for analysis. While these endeavors yielded valuable insights and outcomes, it is imperative to recognize that a more well-balanced and comprehensive dataset has the potential to furnish deeper insights and deliver more precise conclusions.

Despite the limitations of the dataset, it was chosen based on its accessibility and relevance to the research question. The dataset employed, although imperfect, seemed to be the most fitting option for conducting this investigation. Nevertheless, it is vital to acknowledge that the dataset's imperfections may have introduced a certain degree of bias into our findings.

This study carries various constraints. Primarily, it is restrained by the caliber and representativeness of the dataset utilized. Furthermore, the temporal range of the dataset might not capture protracted effects, and the findings of the study may not be universally applicable to all populations. Additionally, the research was predominantly oriented towards scrutinizing correlations, and it was not conducive to establishing causal connections.

With a forward-looking perspective, forthcoming research in this domain should strive to surmount these constraints. Acquiring a more even-handed and diversified dataset, supplemented by a larger sample size, would be invaluable for achieving more resilient analyses and more precise deductions. Furthermore, embarking on longitudinal studies to scrutinize the enduring consequences of COVID-19 on cardiovascular well-being would yield a more all-encompassing comprehension. Additionally, the integration of more cutting-edge machine learning approaches and the contemplation of other pertinent variables, such as

genetic predispositions, could further boost the precision and depth of the analyses.

In summary, notwithstanding the fact that this study has brought to light captivating insights, it is imperative to acknowledge its boundaries and the plausible sway of dataset disproportions. As we persist in grappling with the worldwide repercussions of COVID-19, the perpetuation of research in this arena carries the assurance of divulging additional subtleties and adding to our continually evolving understanding of how the virus affects cardiovascular health.

Acknowledgements

Sincere thanks to Dr Phillip Smith for his unwavering support, foresight, and exemplary management.

Availability of Data and Materials

The data and materials of this study are available upon request and ready to be shared. For further information, please get in touch with the corresponding author, Nidhi Priyadarshini.

Authors and Contributors

Nidhi Priyadarshini designed the study protocol, participated in the data collection and writing of the draft manuscript, oversaw the execution of the study, participated in data analysis, and critically revised the manuscript for intellectual content, participated in statistical analysis and interpretation of results. Dr. Phillip Smith contributed to the study design, provided guidance on the statistical analysis, and reviewed and edited the manuscript for intellectual content.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Zhang, H., Kang, Z., Gong, H., Xu, D., Wang, J., Li, Z., *et al.* (2020) Digestive System Is a Potential Route of COVID-19: An Analysis of Single-Cell Coexpression Pattern of Key Proteins in Viral Entry Process. *Gut*, **69**, 1010-1018. https://doi.org/10.1136/gutjnl-2020-320953

[2] Wu, F., Liu, M., Wang, A., Lu, L., Wang, Q., Gu, C., *et al.* (2020) Evaluating the Association of Clinical Characteristics with Neutralizing Antibody Levels in Patients Who Have Recovered from Mild COVID-19 in Shanghai, China. *JAMA Internal Medicine*, **180**, 1356-1362. https://doi.org/10.1001/jamainternmed.2020.4616

[3] Yu, A., Prasad, S., Akande, A., Murariu, A., Yuan, S., Kathirkamanathan, S., *et al.* (2020) COVID-19 in Canada: A Self-Assessment and Review of Preparedness and Response. *Journal of Global Health*, **10**, Article ID: 0203104. https://doi.org/10.7189/jogh.10.0203104

[4] bin Othman, M.F. and Yau, T.M.S. (2007) Comparison of Different Classification Techniques Using WEKA for Breast Cancer. In: Ibrahim, F., Osman, N.A.A., Usman,

J. and Kadri, N.A., Eds., 3*rd Kuala Lumpur International Conference on Biomedical Engineering* 2006, Springer, 520-523. https://doi.org/10.1007/978-3-540-68017-8_131

[5] Prajwala, T.R. (2015) A Comparative Study on Decision Tree and Random Forest Using R Tool. *International Journal of Advanced Research in Computer Science and Software Engineering*, **5**, 617-621.

[6] Bari Antor, M., Jamil, A.H.M.S., Mamtaz, M., Monirujjaman Khan, M., Aljahdali, S., Kaur, M., *et al.* (2021) A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. *Journal of Healthcare Engineering*, **2021**, Article ID: 9917919. https://doi.org/10.1155/2021/9917919

[7] Schonlau, M. and Zou, R.Y. (2020) The Random Forest Algorithm for Statistical Learning. *The Stata Journal*: *Promoting communications on statistics and Stata*, **20**, 3-29. https://doi.org/10.1177/1536867x20909688

[8] Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T. and Murray, C.J. (2006) Global and Regional Burden of Disease and Risk Factors, 2001: Systematic Analysis of Population Health Data. *The Lancet*, **367**, 1747-1757. https://doi.org/10.1016/s0140-6736(06)68770-9

[9] Grundy, S.M., *et al.* (2018) Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology*, **73**, 2852-2903.

[10] Yancy, C.W., Jessup, M., Bozkurt, B., Butler, J., Casey, D.E., Colvin, M.M., *et al.* (2017) 2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *Circulation*, **136**, e137-e161. https://doi.org/10.1161/cir.0000000000000509

[11] January, C.T., Wann, L.S., Calkins, H., Chen, L.Y., Cigarroa, J.E., Cleveland, J.C., *et al.* (2019) 2019 AHA/ACC/HRS Focused Update of the 2014 AHA/ACC/HRS Guideline for the Management of Patients with Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society in Collaboration with the Society of Thoracic Surgeons. *Circulation*, **140**, e125-e151. https://doi.org/10.1161/cir.0000000000000665

[12] Konstantinides, S.V., Meyer, G., Becattini, C., Bueno, H., Geersing, G., Harjola, V., *et al.* (2019) 2019 ESC Guidelines for the Diagnosis and Management of Acute Pulmonary Embolism Developed in Collaboration with the European Respiratory Society (ERS): The Task Force for the Diagnosis and Management of Acute Pulmonary Embolism of the European Society of Cardiology (ESC). *European Heart Journal*, **41**, 543-603. https://doi.org/10.1093/eurheartj/ehz405

[13] Zeiler, M.D. and Fergus, R. (2014) Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *Computer Vision—ECCV* 2014, Springer, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53

[14] Szegedy, C., Wei Liu,, Yangqing Jia,, Sermanet, P., Reed, S., Anguelov, D., *et al.* (2015) Going Deeper with Convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/cvpr.2015.7298594

[15] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*),

Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[16] Howard, A.G., Zhu, M.L., Chen, B., Kalenichenko, D., Wang, W.J., Weyand, T., Andreetto, M. and Adam, H. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv: 1704.04861.

[17] Li, H., Kadav, A., Durdanovic, I., Samet, H. and Graf, H.P. (2017) Pruning Filters for Efficient ConvNets. *Proceedings of the* 5*th International Conference on Learning Representations* (*ICLR*), Toulon, 24-26 April 2017.

[18] Devlin, J., Chang, M.W., Lee, Ke. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, 2-7 June 2019, 4171-4186.

[19] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., *et al.* (2019) ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 3027-3035. https://doi.org/10.1609/aaai.v33i01.33013027

[20] Guo, T., Fan, Y., Chen, M., Wu, X., Zhang, L., He, T., *et al.* (2020) Cardiovascular Implications of Fatal Outcomes of Patients with Coronavirus Disease 2019 (COVID-19). *JAMA Cardiology*, **5**, 811-818. https://doi.org/10.1001/jamacardio.2020.1017

[21] Azevedo, R.B., Botelho, B.G., Hollanda, J.V.G.d., Ferreira, L.V.L., Junqueira de Andrade, L.Z., Oei, S.S.M.L., *et al.* (2020) COVID-19 and the Cardiovascular System: A Comprehensive Review. *Journal of Human Hypertension*, **35**, 4-11. https://doi.org/10.1038/s41371-020-0387-4

[22] Batlle, D., Soler, M.J., Sparks, M.A., Hiremath, S., South, A.M., Welling, P.A., *et al.* (2020) Acute Kidney Injury in COVID-19: Emerging Evidence of a Distinct Pathophysiology. *Journal of the American Society of Nephrology*, **31**, 1380-1383. https://doi.org/10.1681/asn.2020040419

[23] Johansson, A., Mohamed, M.S., Moulin, T.C. and Schiöth, H.B. (2021) Neurological Manifestations of COVID-19: A Comprehensive Literature Review and Discussion of Mechanisms. *Journal of Neuroimmunology*, **358**, Article ID: 577658. https://doi.org/10.1016/j.jneuroim.2021.577658

[24] Núñez-Gil, I.J., Feltes, G., Viana-Llamas, M.C., Raposeiras-Roubin, S., Romero, R., Alfonso-Rodríguez, E., *et al.* (2023) Post-COVID-19 Symptoms and Heart Disease: Incidence, Prognostic Factors, Outcomes and Vaccination: Results from a Multicenter International Prospective Registry (HOPE 2). *Journal of Clinical Medicine*, **12**, Article 706. https://doi.org/10.3390/jcm12020706

[25] Aghagoli, G., Gallo Marin, B., Soliman, L.B. and Sellke, F.W. (2020) Cardiac Involvement in COVID-19 Patients: Risk Factors, Predictors, and Complications: A Review. *Journal of Cardiac Surgery*, **35**, 1302-1305. https://doi.org/10.1111/jocs.14538

[26] Xie, Y., Xu, E., Bowe, B. and Al-Aly, Z. (2022) Long-Term Cardiovascular Outcomes of Covid-19. *Nature Medicine*, **28**, 583-590. https://doi.org/10.1038/s41591-022-01689-3

[27] Gupta, D., Chauhan, M., Jain, A. and Dewan, S. (2022) Inflammatory Markers as Early Predictors of Disease Severity in COVID-19 Patients Admitted to Intensive Care Units: A Retrospective Observational Analysis. *Indian Journal of Critical Care Medicine*, **26**, 484-488. https://doi.org/10.5005/jp-journals-10071-24171

[28] Passarelli-Araujo, H., Passarelli-Araujo, H., Urbano, M.R. and Pescim, R.R. (2022) Machine Learning and Comorbidity Network Analysis for Hospitalized Patients with COVID-19 in a City in Southern Brazil. *Smart Health*, **26**, Article ID: 100323.

https://doi.org/10.1016/j.smhl.2022.100323

[29] Raman, B., Bluemke, D.A., Lüscher, T.F. and Neubauer, S. (2022) Long COVID: Post-Acute Sequelae of COVID-19 with a Cardiovascular Focus. *European Heart Journal*, **43**, 1157-1172. https://doi.org/10.1093/eurheartj/ehac031

[30] Ambrosino, P., Calcaterra, I.L., Mosella, M., Formisano, R., D'Anna, S.E., Bachetti, T., *et al.* (2022) Endothelial Dysfunction in COVID-19: A Unifying Mechanism and a Potential Therapeutic Target. *Biomedicines*, **10**, Article 812. https://doi.org/10.3390/biomedicines10040812

[31] Sarker, I.H. (2021) Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, **2**, Article No. 160. https://doi.org/10.1007/s42979-021-00592-x

[32] Sarker, I.H. (2022) AI-Based Modeling: Techniques, Applications and Research Issues towards Automation, Intelligent and Smart Systems. *SN Computer Science*, **3**, Article No. 158. https://doi.org/10.1007/s42979-022-01043-x

[33] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) Language Models Are Unsupervised Multitask Learners. Open AI Technical Report.

[34] Rogers, S. and Girolami, M. (2016) A First Course in Machine Learning. 2nd Edition, CRC Press, 397.

[35] Cramer, J.S. (2002) The Origins of Logistic Regression. Tinbergen Institute Discussion Paper No. 02-119/4, 1-16.

[36] Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2012) Foundations of Machine Learning. MIT Press.

[37] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, **26**, 3111-3119.