CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The emergence of the novel coronavirus disease (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has ushered in an unprecedented global health crisis, challenging healthcare systems worldwide. At the time of the writing of this paper, 774 million cases of COVID-19 and 7 million deaths due to COVID-19 have been reported [1,2]. As the pandemic continues to evolve, understanding the clinical course of the disease and developing effective strategies for patient management have become paramount. One critical aspect that demands immediate attention is the timely prediction of intensive care unit (ICU) admission for COVID-19 patients, enabling healthcare professionals to proactively allocate resources, optimise treatment plans, and improve overall patient outcomes [3].

COVID-19 typically presents with a range of symptoms, from mild respiratory distress to severe pneumonia and acute respiratory distress syndrome (ARDS). Common symptoms include fever, cough, shortness of breath, and fatigue, while more severe cases may exhibit tremors, palpitations, and amnesia [4]. Identifying patients at an early stage who are at an increased risk of progressing to severe disease is crucial for allocating limited healthcare resources efficiently and providing timely interventions.

Despite the development and deployment of vaccines, the virus persisted, compounded by the emergence of variants that posed challenges to containment efforts. This unprecedented health crisis wrought profound socio-economic disruptions worldwide, underscoring the necessity for robust global cooperation and continued vigilance to mitigate its impact and chart a course towards recovery.

This chapter aims to present an outline of our project's objectives and problem statement. It lists the key goals we've set out to achieve through our initiative, providing a clear understanding of the purpose and direction of our efforts.

1.1 General Overview

Using machine learning to predict ICU admission due to COVID-19 represents a significant advancement in healthcare management during the pandemic. This approach involves leveraging sophisticated computational algorithms to analyse vast amounts of heterogeneous medical data collected from COVID-19 patients. These data encompass a wide range of variables, including demographic information, pre-existing medical conditions,

symptoms, laboratory findings, imaging results, and vital signs. By harnessing the power of machine learning, these diverse data points are synthesized to identify complex patterns and relationships that may indicate a patient's likelihood of requiring intensive care.

Machine learning models employed in this context are trained on historical patient data, where ICU admission outcomes are known, to learn the intricate associations between various clinical features and the need for critical care. These models utilize a variety of techniques, such as logistic regression, decision trees, support vector machines, or deep learning algorithms, to generate predictions. Through iterative training and validation processes, these models are fine-tuned to optimize their performance and accuracy in predicting ICU admission for COVID-19 patients.

The potential benefits of employing machine learning for ICU prediction in COVID-19 cases are manifold. By providing early identification of patients at higher risk of deteriorating and requiring intensive care, healthcare providers can implement timely interventions, allocate resources more efficiently, and optimize patient care pathways. Moreover, accurate risk stratification can aid in triaging patients, particularly in overwhelmed healthcare systems, where prioritizing treatment for those most in need is crucial. Additionally, insights gleaned from machine learning models can inform clinical decision-making and guide healthcare policies aimed at mitigating the impact of the pandemic.

However, challenges remain in deploying machine learning models for ICU prediction in real-world clinical settings. Ensuring the reliability, interpretability, and generalizability of these models is paramount to their successful integration into healthcare workflows. Furthermore, addressing issues related to data quality, privacy concerns, and model explainability is essential to foster trust among healthcare professionals and ensure ethical deployment. Despite these challenges, leveraging machine learning to predict ICU admission due to COVID-19 holds promise in enhancing patient care, optimizing resource allocation, and ultimately saving lives during this unprecedented global health crisis.

1.2 COVID-19

COVID-19, short for Coronavirus Disease 2019, is a highly contagious respiratory illness caused by the novel coronavirus SARS-CoV-2. The virus first emerged in Wuhan, China, in late 2019 and quickly spread globally, leading the World Health Organization (WHO) to declare it a pandemic in March 2020. COVID-19 primarily spreads through respiratory droplets

when an infected person coughs, sneezes, or talks, making close contact with infected individuals a significant risk factor for transmission. Additionally, the virus can also spread by touching contaminated surfaces and then touching the face. The most common symptoms of COVID-19 include fever, cough, and difficulty breathing, although individuals may experience a wide range of symptoms, from mild to severe. While many people recover from COVID-19 without needing hospitalization, the disease can cause severe respiratory distress, organ failure, and death, particularly in older adults and those with underlying health conditions.

Efforts to control the spread of COVID-19 have involved a combination of public health measures, including widespread testing, contact tracing, quarantine, social distancing, and vaccination campaigns. Multiple vaccines have been developed and authorized for emergency use, demonstrating efficacy in preventing severe illness and reducing transmission rates. Despite vaccination efforts, challenges remain in achieving global vaccine distribution, addressing vaccine hesitancy, and mitigating the impact of emerging variants of the virus. The COVID-19 pandemic has had profound socio-economic impacts, disrupting daily life, straining healthcare systems, and exacerbating disparities in access to healthcare and resources. Thus, the ongoing response to COVID-19 requires continued vigilance, collaboration, and innovation to control the spread of the virus, protect public health, and promote recovery and resilience in affected communities.

1.3 Decision Trees

Decision Trees, a fundamental machine learning technique, have gained widespread recognition for their interpretability and simplicity, making them applicable across various domains, including healthcare. In the context of COVID-19 research, Decision Trees offer unique advantages in understanding and explaining the factors influencing ICU admission [5].

The core principle of Decision Trees involves recursively partitioning the data based on features to create a tree-like structure. Each node in the tree represents a decision or a split, and each leaf node provides the final prediction. This tree structure allows for straightforward interpretation, enabling healthcare professionals to easily grasp the decision-making process. By examining the decision paths within the tree, healthcare professionals can gain a holistic understanding of the patient's condition and the key features driving the predictions [6].

One limitation of decision trees is their tendency to overfit the training data, capturing noise or outliers that may not generalize well to unseen data. To mitigate this issue, techniques such as pruning, which removes parts of the tree that do not contribute significantly to its

predictive accuracy, or ensemble methods like random forests or gradient boosting are often employed. These methods combine multiple decision trees to improve predictive performance while retaining the interpretability of individual trees. Decision trees find applications in various domains, including healthcare, finance, and marketing, where transparent decision-making processes are essential. Their simplicity, interpretability, and ability to handle both numerical and categorical data make decision trees a valuable tool in the machine learning toolkit.

1.4 Problem Statement

The problem statement revolves around the need to enhance ICU admission prediction specifically tailored for COVID-19 patients using decision trees. Despite advancements in medical technology and treatment protocols, accurately identifying which COVID-19 patients will require intensive care remains a challenge.

Several challenges may arise when enhancing ICU admission prediction for COVID-19 patients using decision trees. Firstly, ensuring the availability and quality of data is crucial. COVID-19 patient data may be heterogeneous and collected from various sources, leading to inconsistencies and missing values. Additionally, variations in testing protocols and documentation practices across healthcare facilities can affect data completeness and accuracy, potentially impacting the reliability of the predictive model.

Interpretability of the decision tree model is another concern. While decision trees are inherently interpretable, complex or deep trees may become difficult to understand, hindering their utility in clinical decision-making. Ensuring that the decision tree remains interpretable while capturing the nuanced relationships between patient features and ICU admission is essential for gaining trust and acceptance from healthcare professionals.

Furthermore, the generalizability of the decision tree model is critical for its applicability across different patient populations and healthcare settings. COVID-19 patient demographics, disease severity, and treatment protocols may vary geographically and temporally, necessitating robust validation strategies to assess the model's performance across diverse cohorts. Failure to address these issues may result in poor model performance or biased predictions, limiting its effectiveness in real-world clinical practice.

Moreover, the dynamic nature of the COVID-19 pandemic poses additional challenges. Emerging variants of the virus, changes in clinical guidelines, and fluctuations in healthcare

resource availability may impact the predictive performance of the decision tree model over time. Continuous monitoring and adaptation of the model are essential to ensure its relevance and accuracy amidst evolving circumstances.

Overall, addressing these challenges is crucial to develop a reliable and effective decision tree-based approach for ICU admission prediction in COVID-19 patients, ultimately contributing to improved patient care and healthcare resource management during the ongoing pandemic.

1.5 Objective

The primary objective of this project is to enhance critical resource allocation within healthcare systems, particularly in the context of managing COVID-19 patient care. By developing and implementing a robust predictive model for ICU admission using decision trees, the aim is to ensure the efficient allocation of limited healthcare resources and enable timely interventions for patients at higher risk of requiring intensive care. Through accurate risk stratification and early identification of high-risk COVID-19 cases, healthcare providers can optimize the allocation of ICU beds, ventilators, medical staff, and other critical resources, ultimately improving patient outcomes and reducing healthcare system strain.

In pursuit of this objective, the project seeks to introduce and evaluate the novel Omega index for feature selection and assess its impact on reducing the time required for ICU admission prediction. The Omega index, as a novel metric for evaluating the importance of features in predictive models, holds promise in enhancing the efficiency and effectiveness of feature selection processes. By integrating the Omega index into the decision tree-based predictive model, the project aims to streamline the feature selection process and improve the model's performance in predicting ICU admission for COVID-19 patients. This exploration of the integration of the Omega index with decision trees represents a novel and practical methodology that has the potential to advance the field of predictive analytics in healthcare, particularly in critical domains such as COVID-19 patient care.

1.6 Motivation

Our research is driven by the imperative to enhance healthcare practices and address the ongoing challenges in patient management. Despite the gradual transition away from the peak of the COVID-19 pandemic, the need for effective predictive tools remains critical. Our study specifically focuses on ICU admission prediction for COVID-19 patients, recognizing the importance of timely interventions and resource allocation in optimizing patient outcomes. By

leveraging advanced techniques such as decision trees and innovative metrics like the Omega index, we aim to empower healthcare providers with reliable tools to navigate the complexities of patient care in today's healthcare landscape.

In the absence of the acute crisis atmosphere that characterized the peak of the pandemic, our research underscores the enduring importance of proactive strategies in healthcare. The lessons learned from the COVID-19 era have highlighted the value of predictive analytics in guiding clinical decision-making and improving patient outcomes. Our endeavour in 2024 reflects a commitment to building upon these lessons, leveraging cutting-edge methodologies to enhance healthcare practices and address the evolving needs of patient management. Through our research, we strive to contribute to a more resilient and responsive healthcare system capable of effectively managing patient care in both routine and challenging circumstances.

1.7 Applications

The project on enhancing ICU admission prediction for COVID-19 patients using decision trees and the Omega index has several potential applications:

- **Clinical Decision Support:** The predictive model can serve as a valuable tool for healthcare providers in making timely and informed decisions regarding ICU admission for COVID-19 patients, facilitating proactive patient management.
- **Resource Allocation:** By accurately predicting ICU admission, healthcare systems can optimize the allocation of critical resources such as ICU beds, ventilators, and medical staff, ensuring that resources are directed to those who need them most.
- **Triage and Prioritization:** The predictive model can aid in triaging COVID-19 patients based on their risk of requiring intensive care, allowing healthcare providers to prioritize treatment for high-risk individuals and allocate resources accordingly.
- **Healthcare Planning and Preparedness:** Insights gained from the predictive model can inform healthcare planning and preparedness efforts, helping healthcare systems anticipate and respond to surges in COVID-19 cases more effectively.
- **Research and Public Health Surveillance:** The project's findings can contribute to research efforts aimed at understanding the factors influencing ICU admission for COVID-19 patients and informing public health surveillance strategies to monitor and mitigate the spread of the virus.

- **Clinical Trials and Drug Development:** The predictive model can be used to identify high-risk patient populations for clinical trials and drug development efforts aimed at improving outcomes for COVID-19 patients requiring intensive care.
- **Healthcare Policy and Decision-Making:** Insights from the project can inform healthcare policy and decision-making processes, guiding efforts to improve patient outcomes and allocate resources efficiently in response to the ongoing COVID-19 pandemic and future healthcare crises.

1.8 Contribution

Within the scope of this project, we introduce a novel index known as the Omega index which has been designed to streamline the process of feature selection. The primary objective of this index is to select the relevant features, contributing to the optimization of the overall prediction model. By carefully identifying and incorporating key variables through this index, we aim to significantly reduce the time required for ICU admission prediction. The integration of the Omega index with Decision trees represents a novel and practical methodology for advancing the field of predictive analytics in healthcare, particularly in the critical domain of COVID-19 patient care.

1.9 Organisation

This report details the development of a COVID-19 diagnosis system using decision trees.

Chapter 1: Introduction provides a general overview of COVID-19, decision trees, and the project's purpose. It outlines the problem statement, objectives, motivations, and potential applications of this system.

Chapter 2: Literature Survey delves into existing research on decision tree applications in diagnosing COVID-19 and related diseases. This chapter identifies relevant studies and evaluates their methodologies and findings.

Chapter 3: System Requirements Specification details the hardware, software, functional (specific tasks the system performs), and non-functional (characteristics like usability and security) requirements for building and deploying the system.

Chapter 4: System Design explores design considerations, provides an overview of the system's architecture, and discusses existing work in this field. It outlines the model development procedure, outlining the steps involved in creating the decision tree model.

Chapter 5: System Implementation dives into the specifics of data acquisition and preprocessing. It then explains the chosen evaluation metric (Omega Index) and details the implementation of the decision tree model itself.

Chapter 6: Results and Analysis presents the system's performance metrics and analyses the effectiveness of the decision tree model in diagnosing COVID-19.

Chapter 7: Conclusion and Future Scope summarizes the project's achievements, highlights key findings, and discusses potential areas for future research and development of the system.

CHAPTER 2: LITERATURE SURVEY

Various studies have explored machine learning and deep learning techniques for managing and analysing COVID-19. Cosimo Magazzino et al. correlated vaccination data with a significant reduction in COVID-19 fatality rates using an ANN. Shrikant Tiwari et al. discussed widespread ML usage, including LR, RF, kNN, and SVM. Yogesh H. Bhosale's reviews focused on deep learning applications in COVID-19 diagnosis, noting changes in lesion patterns. Ameer Sardar Kwekha-Rashid et al. analysed diverse ML models, with Logistic Regression being frequently applied. Kirti Raj Bhatele et al. conducted a systematic review on deep learning-based COVID-19 detection using chest X-rays and CT scans. Arash Heidar et al. explored ML for COVID-19 outbreak management, emphasising the diverse application areas. E. Gothai et al. used LR and SVR for COVID-19 trend prediction, highlighting the importance of model validation.

The papers collectively highlight the diverse applications of machine learning and deep learning in managing and analysing COVID-19, demonstrating their efficacy in vaccination impact assessment, severity prediction, diagnosis, and outbreak management.

Cosimo Magazzino et al. [7] presented a machine learning algorithm aimed at study involving utilising an ANN to analyse data from a sample of N = 192 countries worldwide. The dataset included information on vaccines, confirmed COVID-19 cases, and fatality rates, covering the period from March to May 2021. The study's findings revealed a significant correlation between vaccination plans, turnout in vaccination campaign participation, doses administered, and a sudden reduction in the fatality rate of COVID-19. This reduction was particularly notable precisely at the point where the cut effect was triggered in the Neural Network (NN).

Shrikant Tiwari et al.[8] conducted a comprehensive study that delves into various machine learning algorithms employed in the analysis of Covid-19 cases. Among the algorithms explored are MLP, kNN, SVM, RF, ANFIS, NB, LR, LIR methods. The review highlights the widespread utilisation of machine learning in various aspects of COVID-19 analysis, including forecasting, detection, identification, and screening. This application has been particularly prominent since December 2019 and continues to the present day. Notably, machine learning methods such as LR, RF, kNN, and SVM have played crucial roles in addressing the challenges posed by COVID-19.

Yogesh H. Bhosale et al.[9] conducted a systematic review study that primarily focused on leveraging deep learning techniques for the diagnosis of Covid-19. The applied methods included CNN and RNN. The review incorporated data from 35 different datasets, offering a comprehensive analysis of deep learning's application in diagnosing Covid-19. The paper contextualises this within the broader landscape of ML applications in COVID-19 management, covering aspects like forecasting, detection, identification, and screening. Traditional ML methods such as LR, RF, kNN, and SVM were discussed for their contributions to addressing COVID-19 challenges. In addition to highlighting recent developments in deep learning techniques for classifying COVID-19 based on diverse lung and chest imaging modalities. However, the study recognized limitations in subsequent X-ray, CT, and ultrasound studies for COVID-19 classification. Notably, the review observed distinct changes in pulmonary lesion patterns, including Ground Glass Opacity (GGO) and high mortality rates during the early phases of COVID-19.

Yogesh H. Bhosale et al.[10] conducted a meta-analysis and systematic review that focused on the applications of deep neural networks and machine learning techniques in the diagnosis, detection, classification, and segmentation of COVID-19, particularly within the realms of bio-medical imaging (X-ray, CT, ultrasound, ECG), and genome sequences. The review incorporated data from 35 different datasets, providing a comprehensive analysis of the utilisation of deep neural networks in the context of COVID-19. The paper identified a range of concerns and challenges associated with this approach, including the absence of a globally certified dataset, issues related to security, regulatory requirements, and the need for transparent information. Furthermore, the study emphasised the importance of addressing differences in illnesses between Covid-19 and non-Covid-19 cases, contributing valuable insights to the ongoing discourse on the application of advanced technologies in the field of COVID-19 diagnosis and management.

Yibai Xiong et al.[11] focused on comparing different machine learning techniques for predicting COVID-19 severity. The utilised techniques include LR, SVM, and RF. Unfortunately, the datasets generated and/or analysed during the study are not publicly available due to the sensitivity of the patients' personal information. Throughout the investigation, several concerns and challenges were identified, encompassing the absence of a globally certified dataset, issues related to the security and regulatory aspects of handling patient information, and the need for transparent information. Furthermore, the study acknowledges variations in illnesses between COVID-19 and non-COVID-19 cases,

highlighting the complexity of predicting severity accurately. Notably, the research suggests that Random Forest (RF) may serve as a valuable prioritisation tool for predicting COVID-19 severity, showcasing favourable overall performance upon admission.

N. Rochmawati et al. [12] conducted the research that focuses on employing DT to assess the severity of COVID-19 symptoms. The Weka application was utilised for the research, and the dataset used in the study was sourced from the public dataset available on the Kaggle website. The study specifically explores the application of J48 and the Hoeffding Tree to establish clear rules for classifying individuals into categories of mild, moderate, severe, or non-COVID. Despite a less significant difference between J48 and the Hoeffding Tree, the study slightly favours J48 in terms of accuracy, precision, and recall. From a tree-view perspective, the Hoeffding Tree is noted for its simplicity, exhibiting fewer nodes compared to J48. This research contributes insights into the decision-making process for categorising COVID-19 symptom severity using decision tree methodologies.

Ameer Sardar Kwekha-Rashid et al.[13] conducted research on various machine learning applications employed for the analysis of COVID-19 cases. The utilised models include MobileNet, DenseNet, Xception, ResNet, InceptionV3, InceptionResNetV2, VGGNet, and NASNet. The study extensively utilised 152 datasets of COVID-19 patients and 500 chest CT scans for the analysis. The application of different machine learning models in COVID-19 case analysis signifies a multidimensional approach to understanding and managing the disease. Interestingly, Logistic Regression emerges as the most frequently applied machine-learning algorithm, featuring in five out of fourteen papers. Following closely, ANN and CNN rank second and third, appearing in three and two papers, respectively. The study also notes the application of other algorithms, such as Linear Regression, K-Means, KNN, and NB, in the specific context of production lines. This comprehensive analysis sheds light on the diverse applications of machine learning in addressing the challenges posed by COVID-19 cases.

Kirti Raj Bhatele et al.[14] focused on machine and deep learning-based approaches for detecting COVID-19, utilising chest X-rays and CT scans. The primary deep learning model employed in this study is CNN. For the analysis, the review utilised both a local CT scans dataset and a global chest X-ray dataset. The systematic review aims to provide a comprehensive analysis of the state-of-the-art methodologies in the realm of deep and machine learning for the detection of COVID-19. By leveraging chest X-rays and CT scans, these

approaches contribute valuable insights to the ongoing efforts in developing effective diagnostic tools for identifying COVID-19 cases.

Arash Heidar et al.[15] explores the use of machine learning techniques for managing the COVID-19 outbreak. The applied machine learning models include Support Vector Machine, Statistical Neural Network, Decision Tree, and Feed-Forward Neural Network. The study, conducted under the name ML-COVID-19, encompasses a variety of attributes related to machine learning applications for COVID-19, covering advantages, challenges, datasets, dataset sizes, applications, privacy considerations, and Transfer Learning (TL). This comprehensive analysis provides insights into the strengths and limitations of different machine learning approaches in the context of managing the COVID-19 outbreak. Additionally, the survey conducted as part of the study reveals the distribution of applications, indicating that 20.4% of them focus on categorising imaging techniques, 16% on monitoring, and 14% on prediction and forecasting. This information sheds light on the diverse areas where machine learning applications have been applied for effective outbreak management during the COVID-19 pandemic.

E. Gothai et al. [16] did a study that employed Linear Regression and Support Vector Machine as machine learning approaches for predicting the growth and trend of COVID-19. The study utilised a serious forecasting method and accessed COVID-related data through the COVID open access provided by the WHO. Specifically, LR and SVR were among the algorithms employed. However, the study noted that LR and SVR yielded fewer confirmed cases compared to the proposed time series Holt's Winter model. Consequently, the predictions made by these algorithms may vary in accuracy. To enhance the robustness of the evaluation, the study emphasised the importance of comparing these predictions with actual data on the number of infected cases, highlighting the need for ongoing refinement and validation of machine learning models in the context of COVID-19 prediction.

TABLE 1.1: Research papers examined as a part of the Literature Survey

Sl. No.	Author and Year	Advantages	Disadvantages	Results
1	Sunanda Das et al. [17] 2023	Leverages powerful feature extraction, Improves accuracy with Transfer Learning, Interpretability with Random Forest, Robust	Computational cost, Data dependence, Black box nature of CNN, Complex	It attains accuracy, precision, recall, and F1-score values of 91.67%, 93.64%, 91.67%, 91.47% for character recognition, and 97.33%, 97.89%, 97.33%, 97.37% for digit recognition, respectively.
2	Hakim El Massari et al. [18] 2022	Machine learning techniques offer a simpler and potentially faster approach. Ontology-based methods can be more powerful.	Machine learning techniques might not capture the richness of knowledge in ontologies. Ontology-based methods require more upfront investment in ontology development.	Ontology classifier stands out as the best performer, boasting a high accuracy of 77.5%. Following closely are the SVM algorithms with 77.3%, and logistic regression with 77.2%.
3	Gajendra Sharma et al. [19] 2022	Random forests are generally less prone to overfitting compared to single decision trees. Feature importance can be extracted from random forests, offering some understanding of which image features contribute most to the classification.	The study focuses on a single model (random forest) and a specific dataset (6000 digits). Generalizability to other classification tasks or larger datasets might be limited. They can also become computationally expensive for very large datasets.	The testing accuracy impressively reaches 99% when utilising 50 decision trees.
4	Tahia Tazin et al. [20] 2021	Training CNNs on large datasets containing stroke and non-stroke images can lead to more accurate and automated detection of	The accuracy of the model heavily relies on the quality and completeness of the data in the dataset. Inconsistent labelling, missing data, or biases in	Achieved a remarkable classification accuracy of 96 percent.

		primary ischemic stroke. Faster and more accurate stroke diagnosis can lead to earlier intervention.	data collection can affect model performance.	
5	Mostafa Shanbehzadeh et al [21] 2022	ANNs have the capability to learn complex patterns from data, potentially leading to accurate identification of COVID-19 cases based on the medical records used.	The study uses data from only 250 confirmed COVID-19 cases and 150 negative cases. This might be insufficient for an ANN to learn robust patterns for reliable diagnosis, potentially leading to overfitting. It can be challenging to understand how ANNs arrive at their decisions which can hinder trusting and validating the model's recommendations.	ROC of 0.982, a sensitivity of 96.4%, a specificity of 90.6%, and an accuracy of 94% was achieved.
6	Azhari A. Elhag et al. [22] 2021	This study suggests MLPs might achieve better classification accuracy compared to simpler models like Logistic Regression. This could be crucial in medical applications where accurate diagnosis is essential. MLPs, can learn complex non-linear relationships between input data.	The data covers a specific timeframe and the model's performance on data from different time periods or geographical locations might be lower. The study only compares MLP with LRM. It's unclear how MLP performs against other potentially more suitable models for medical diagnosis.	The results demonstrated that the Multilayer Perceptron (MLP) outperformed the Logistic Regression Model (LRM), with a higher classification accuracy rate of 85.6% compared to 80.8%.
7	Govardhan Jain et al. [23] 2020	The study suggests deep learning has the potential to be a highly accurate tool for diagnosing	The model's performance relies heavily on the quality and representativeness	The first stage model shows accuracy of 93.01%. The second stage model to detect the presence of

		COVID-19 using chest X-rays. Deep learning models can be applied to large datasets of chest X-rays, potentially leading to improved diagnosis accuracy over time as the model learns from more data	of the datasets used (Cohen and Kaggle). Biases or limitations in these datasets could affect the model's generalizability. Understanding how the model identifies COVID-19 cases remains unclear, limiting transparency in medical settings.	COVID-19 shows an exceptional performance with an accuracy of 97.22%.
8	Mostafa Shanbehzadeh et al. [24] 2021	Decision trees are known for their clear and interpretable structure. This transparency can be valuable for in understanding the model's reasoning. Training decision trees is often computationally less expensive compared to other machine learning models like deep learning.	Decision trees can be prone to overfitting, especially with smaller datasets. Decision trees primarily capture linear relationships between features. Complex non-linear relationships between various factors, which might be present in medical data, could be missed by the model.	The results indicated that the J-48, with the accuracy= 0.85, F-Score= 0.85, ROC= 0.926, and PRC= 0.93, had the best performance for diagnosing COVID-19.
9	Khadijeh-Moulaei et al. [26] 2022	Random Forests are generally less prone to overfitting compared to single decision trees, potentially leading to more robust performance on unseen data. They can capture complex non-linear relationships between features in the data.	The performance of the RF used in this study, heavily relies on the quality, completeness, and representativeness of the data used for training. Real-world application requires validation on external datasets from different hospitals to ensure generalizability.	The RF model demonstrated an accuracy of 95.03%, sensitivity of 90.70%, precision of 94.23%, specificity of 95.10%, and a receiver operating characteristic (ROC) of 99.02%.

CHAPTER 3: SYSTEM REQUIREMENTS SPECIFICATION

The system requirements specification aims to define the program needs and interfaces to enhance usability, playing a pivotal role in shaping the software's user experience. This chapter comprehensively addresses the project's needs and outlines both functional and non-functional requirements, offering insights to assist new users in navigating the software. The Software Requirements Specification (SRS) document further elaborates on the functional, behavioural, and data requirements, serving as a blueprint for system behaviour. It provides a detailed explanation of the system's anticipated behaviour before its creation, emphasizing the importance of thorough requirement analysis. Requirement analysis ensures that requirements are quantifiable, testable, and aligned with business opportunities or needs, laying the groundwork for successful system design and development.

3.1 Hardware Requirements

- **Processor**: A multi-core processor such as Intel Core i7 or AMD Ryzen 7, capable of handling parallel processing, is recommended for efficient computation.
- **Memory (RAM)**: A minimum of 8 GB RAM, such as DDR4 3200MHz, is required to handle large datasets effectively. For larger datasets or more complex algorithms, 16 GB or more may be necessary.
- **Storage**: Adequate storage space, preferably on solid-state drives (SSDs) such as Samsung 970 EVO or Crucial MX500, is essential for storing datasets, software libraries, and model files.
- **Graphics Processing Unit (GPU)**: While not strictly necessary, a dedicated GPU such as NVIDIA GeForce RTX 3060 or AMD Radeon RX 5700 XT can significantly accelerate computations, especially for large-scale data processing tasks.

3.2 Software Requirements

- **Programming Languages**: Python 3.8 or higher.
- **Machine Learning Libraries**: Libraries such as Scikit-learn which provide implementations of decision tree algorithms and other machine learning models.
- **Integrated Development Environment (IDE)**: IDEs such as Jupyter Notebook, which allows for interactive data exploration and visualization.
- **Data Processing Tools**: Tools like Pandas are essential for cleaning and transforming datasets before model training.

- **Visualization Tools**: Visualization libraries such as Matplotlib, Seaborn, or Plotly enable the creation of data visualizations to explore data distributions, model performance metrics, and prediction results.

3.3 Functional Requirements

A functional requirement specifies what the system should do, describing specific functions or features that the system must perform to satisfy user needs. These requirements define the behaviour of the system in response to inputs and describe the system's functionality from the user's perspective.

- **Data Collection**: The system shall collect diverse patient data, including demographic information, medical history, symptoms and laboratory results from multiple sources such as electronic health records (EHRs), hospital databases, and public health repositories.
- **Data Preprocessing**: The system shall preprocess the collected data to handle missing values, outliers, and inconsistencies, as well as perform feature engineering and data transformation tasks to prepare the data for model training.
- **Model Development**: The system shall develop decision tree models using the pre-processed data to predict the likelihood of ICU admission for COVID-19 patients based on their clinical characteristics and other relevant factors.
- **Model Evaluation**: The system shall evaluate the performance of the decision tree models using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score, using techniques such as cross-validation to ensure robustness and generalizability.
- **Real-time Prediction**: The system shall provide functionality for real-time prediction of ICU admission likelihood for new COVID-19 patients based on their input data, allowing for timely interventions and resource allocation.
- **Model Interpretability**: The system shall provide functionality to interpret the decision tree models and visualize important features and decision pathways, enabling healthcare professionals to understand and trust the model's predictions.
- **Integration with Existing Systems**: The system shall integrate seamlessly with existing healthcare IT systems and workflows, allowing for easy deployment and adoption within healthcare institutions.

3.4 Non-functional Requirements

A non-functional requirement specifies how the system should perform, focusing on quality attributes or constraints that the system must adhere to. These requirements describe system

characteristics such as performance, reliability, security, usability, and scalability. Non-functional requirements are not directly related to specific functions or features but rather describe the overall behaviour, characteristics, and constraints of the system.

- **Performance**: The system shall provide predictions within a specified response time, ensuring timely decision-making and interventions for patient care. For example, predictions should be generated within seconds or minutes to facilitate real-time clinical decision support.
- **Accuracy**: The prediction models shall achieve a minimum accuracy rate, such as 90%, in identifying COVID-19 patients requiring ICU admission.
- **Scalability**: The system shall be scalable to handle increasing volumes of patient data and user interactions without sacrificing performance. It should be able to accommodate growing datasets and user loads, such as during periods of high COVID-19 case numbers.
- **Reliability**: The system shall be highly reliable, with minimal downtime and errors, to ensure continuous availability for healthcare professionals relying on ICU admission predictions for patient care decisions.
- **Interpretability**: The prediction models shall be interpretable, with clear explanations of how predictions are made and the factors influencing them. Healthcare professionals should be able to understand and trust the model's predictions to support informed decision-making.

CHAPTER 4: SYSTEM DESIGN

In the context of this project, system design plays a pivotal role in shaping the architecture and functionality of the predictive model. This phase serves as the blueprint for developing a robust and reliable system capable of accurately predicting the likelihood of ICU admission for COVID-19 patients based on their clinical characteristics. The system design process involves translating the project's requirements and objectives into a detailed plan that guides the implementation, testing, and deployment of the predictive model.

4.1 Design Considerations

The system design for this project encompasses various aspects, including data collection, preprocessing, model development, evaluation, and deployment. It involves defining the structure and organization of the predictive model, specifying the algorithms, features, and parameters necessary for accurate prediction. Additionally, system design considerations include scalability, reliability, performance, security, and usability to ensure that the resulting predictive model meets the needs of healthcare professionals and stakeholders effectively. Throughout the system design phase, careful consideration is given to factors such as the size and complexity of the dataset, computational resources required, model interpretability, and integration with existing healthcare systems.

4.2 Overview of System Design

In this system design brief, we outline the implementation of a novel index called the Omega index, aimed at enhancing feature selection within predictive analytics for healthcare, specifically focusing on ICU admission prediction in the context of COVID-19 patient care. The Omega index serves as a streamlined approach to identify and integrate relevant variables crucial for optimizing prediction models. By leveraging Decision trees in conjunction with the Omega index, we anticipate a substantial reduction in the time required for ICU admission prediction, thus enabling more efficient and accurate patient care management. This innovative methodology promises to advance the field of predictive analytics in healthcare, offering practical solutions to the challenges posed by the critical domain of COVID-19 patient care.

4.3 Existing Work

Existing work in the field of predictive analytics for healthcare, particularly in ICU admission prediction, encompasses a variety of methodologies and approaches. Traditional methods often rely on statistical techniques such as logistic regression or machine learning

algorithms like support vector machines and random forests to predict ICU admissions based on patient data including vital signs, lab results, and comorbidities. Recent advancements in predictive analytics have seen the incorporation of deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which demonstrate improved performance in capturing complex patterns and temporal dependencies within patient data. Moreover, feature selection techniques play a crucial role in enhancing the performance and interpretability of predictive models. Existing methods include filter methods, wrapper methods, and embedded methods, each with its advantages and limitations. However, despite these advancements, challenges remain in accurately predicting ICU admissions, especially in dynamic healthcare environments such as those seen during the COVID-19 pandemic. These challenges include data scarcity, class imbalance, model interpretability, and real-time implementation in clinical settings. In response to these challenges, researchers have proposed novel approaches such as ensemble learning, feature engineering, and hybrid models combining multiple algorithms to improve prediction accuracy and robustness.

Overall, existing work in ICU admission prediction demonstrates a diverse range of methodologies and approaches, reflecting the ongoing efforts to address the complex and evolving nature of healthcare data analysis in critical care settings.

4.4 Outline of the Model Development Procedure

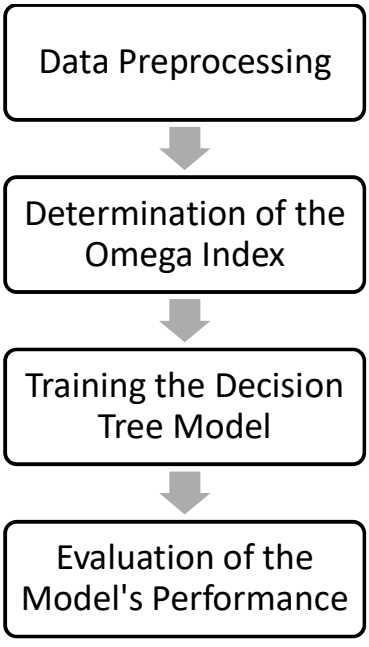

Fig. 4.1: Model development methodology

The model development methodology has been summarised in Fig 4.1. The steps have been elaborated upon in detail below.

1. **Data Preprocessing**: Begin the data preprocessing phase by thoroughly cleaning and organising the dataset. Address any missing values or outliers that might impede the analysis. Optimise data for model training by ensuring it is appropriately formatted to meet the requirements of the machine learning model.
2. **Determination of the Omega Index**: Compute the Omega Index for each feature within the dataset. This involves evaluating the significance of each feature and assigning them values accordingly. Then, rank the features based on the computed Omega Index values, to determine their respective contributions to the analysis.
3. **Training the Decision Tree Model**: With the ranked features in hand, proceed to train the Decision Tree model by iteratively varying the number of features selected from the previously obtained ranking. Assess the model's performance in each iteration and determine the optimal number of features needed for prediction by observing the model's performance until it reaches a point where evaluation metrics show saturation.
4. **Evaluation of the Model's Performance**: Now, with the optimal number of features, evaluate the model by utilising a distinct dataset that was not part of the training process to gauge how well the model generalises to new, unseen data. Calculate key metrics such as accuracy, precision, recall, and others, providing a comprehensive insight into the effectiveness of the Decision Tree model in addressing the specific problem at hand. This thorough evaluation ensures a robust understanding of the model's capabilities and limitations.

CHAPTER 5: SYSTEM IMPLEMENTATION

The system implementation phase translates the system design into a working system. Data collection methods are put into practice and the raw data is cleaned and prepared for the chosen model. The model itself is built using the specified algorithms, features, and parameters. During implementation, we prioritize scalability, reliability, performance, security, and user-friendliness. This ensures the final system effectively addresses the needs of healthcare professionals.

5.1 Dataset and preprocessing

The dataset contains anonymized data of 385 patients from Hospital Sírio-Libanês, São Paulo and Brasilia [27]. It encompasses the patient's demographic details (03), the patient's medical history (via grouped diseases) (09), blood results (36 parameters) and vital signs data (06). All data were anonymized following the best international practices and recommendations and comprises critical patient information, delineated across various categories and have been cleaned and scaled by column according to Min Max Scaler to fit between -1 and 1. The data was obtained and aggregated by windows in chronological order and is depicted in Table 5.1.

TABLE 5.1: Description of time windows

Window	Description
0 - 2	From 0 to 2 hours of the admission
2 - 4	From 2 to 4 hours of the admission
4 - 6	From 4 to 6 hours of the admission
6 - 12	From 6 to 12 hours of the admission
Above 12	Above 12 hours from admission

In total there are 54 features (excluding the Patient identifier and the target variable ICU), expanded when pertinent to the mean, median, max, min, diff (difference) and relative diff (relative difference). The features have been tabulated in Table 5.2.

$$diff = max - min \tag{1}$$

$$relative\ diff = \frac{diff}{median} \tag{2}$$

The target variable ICU, indicates whether the patient has been admitted to the ICU during a specified time window. If the ICU variable has a value of 1(indicating admission to the ICU) then, the readings corresponding to that time window and subsequent time windows becomes infeasible for modelling as per the warning issued by the authors of the dataset.

TABLE 5.2: Features in the dataset

Sl. No.	Feature	Sl. No.	Feature
1	PATIENT_VISIT_IDENTIFIER	29	LEUKOCYTES
2	AGE_ABOVE65	30	LINFOCITOS
3	AGE_PERCENTIL	31	NEUTROPHILES
4	GENDER	32	P02_ARTERIAL
5	DISEASE GROUPING 1	33	P02_VENOUS
6	DISEASE GROUPING 2	34	PC02_ARTERIAL
7	DISEASE GROUPING 3	35	PC02_VENOUS
8	DISEASE GROUPING 4	36	PCR
9	DISEASE GROUPING 5	37	PH_ARTERIAL
10	DISEASE GROUPING 6	38	PH_VENOUS
11	HTN	39	PLATELETS
12	IMMUNOCOMPROMISED	40	POTASSIUM
13	OTHER	41	SAT02_ARTERIAL
14	BE_ARTERIAL	42	SAT02_VENOUS
15	BE_VENOUS	43	SODIUM
16	BIC_ARTERIAL	44	TGO
17	BIC_VENOUS	45	TGP
18	BILLIRUBIN	46	TTPA
19	BLAST	47	UREA
20	CALCIUM	48	DIMER
21	CREATININ	49	BLOODPRESSURE_DIASTOLIC
22	FFA	50	BLOODPRESSURE_SISTOLIC
23	GGT	51	HEART_RATE
24	GLUCOSE	52	RESPIRATORY_RATE
25	HEMATOCRITE	53	TEMPERATURE
26	HEMOGLOBIN	54	OXYGEN_SATURATION

27	INR	55	WINDOW
28	LACTATE	56	ICU

The dataset came with the following warning: “Beware NOT to use the data when the target variable is present, as it is unknown the order of the event (maybe the target event happened before the results were obtained)”.

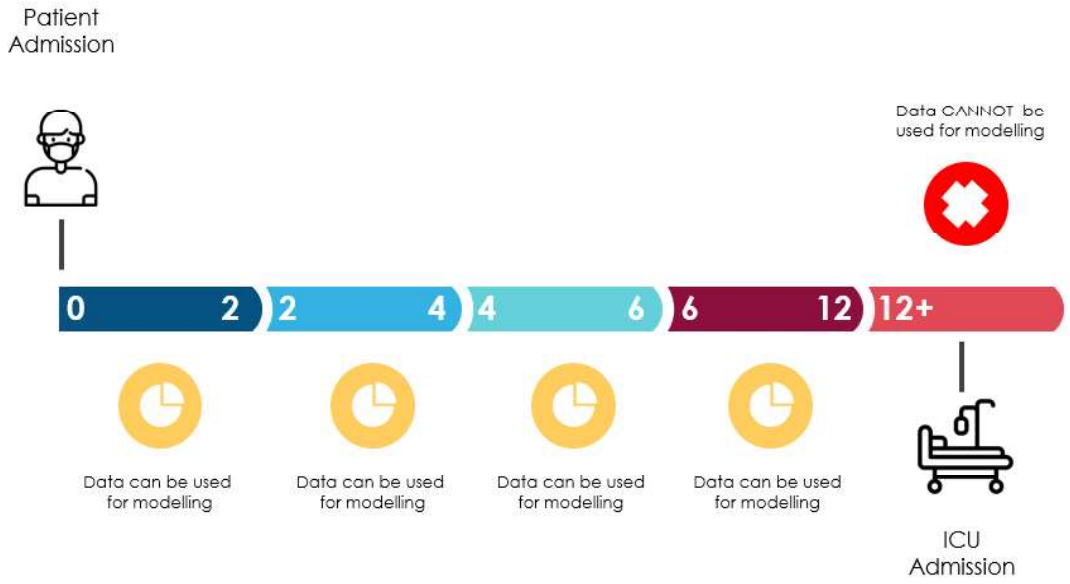

Fig. 5.1: Diagrammatic representation of data that can be used for modelling

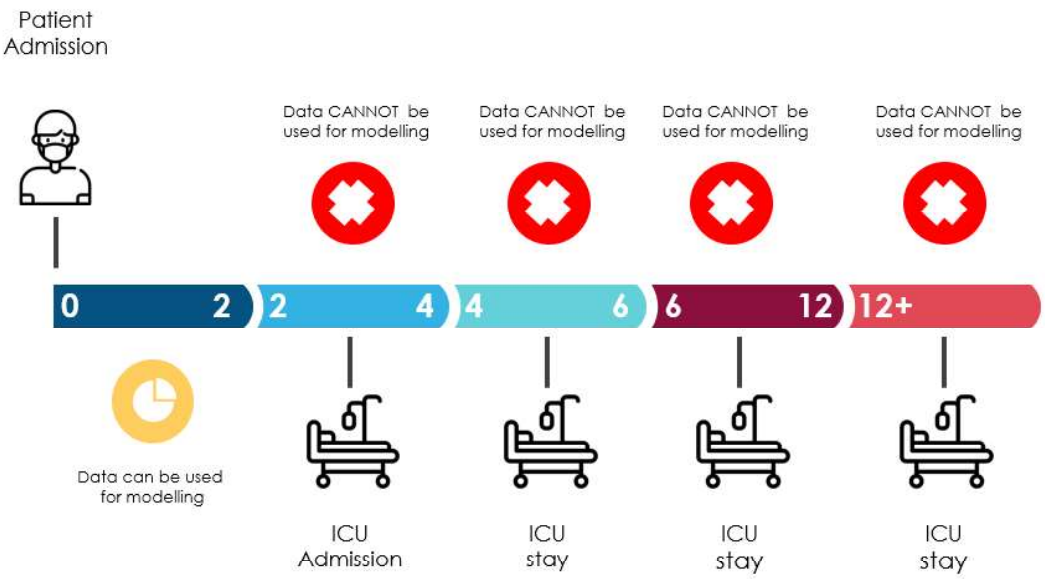

Fig. 5.2: Diagrammatic representation of data that cannot be used for modelling

The data had to be used for modelling in accordance with the details depicted in Fig. 5.1 and Fig 5.2. Consequently, the data had to be processed so as to make it possible to use it for machine learning tasks.

The data set had 5 records per patient so as to correspond to the five windows specified earlier. We aggregated the data to one record per patient to ensure that it would make data processing easier. The target variable, ICU, indicated whether or not the patient had been admitted to the ICU. After the data aggregation process, we ended up with a single record corresponding to each patient with the individual features having an array consisting of 5 elements. Aggregating data at the patient visit level enables a more detailed analysis, capturing specific information for each visit.

We excluded entries from consideration if all five elements in the array associated with the ICU target variable were equal to 1. This exclusion was necessary for data modelling due to the earlier warning as the presence of all 1s indicates that none of the recorded values are appropriate for inclusion in the data modelling process.

We then modified the dataset to represent the last observed value for various features before a patient is admitted to the ICU. This step enhances the dataset's informativeness for subsequent predictive modelling. We transformed the arrays in the 'ICU' field into binary values, assigning '1' if at least one '1' was present in the array, and '0' otherwise. This stage

assists in establishing whether the patient was eventually admitted to the ICU or not. At this point, a compromise is made wherein we sacrifice the precise timing of the patient's entry into the ICU by simplifying it to a binary representation indicating whether the patient was admitted to the ICU or not.

This simplification is acceptable for several reasons. First, in certain analytical or predictive modelling scenarios, the exact timing of when a patient enters the ICU might be less critical than the binary distinction of whether they were admitted or not. Second, transforming the data into a binary outcome can facilitate the development of more straightforward and interpretable models, which is often desirable in healthcare settings. Lastly, by focusing on the broader admission status, the dataset becomes more amenable to various statistical analyses and machine learning techniques, allowing for effective pattern recognition and prediction without the need for precise temporal information.

5.2 Omega Index

For the sake of feature extraction, we planned on using a previously determined index called the alpha index, with a few enhancements of our own. We have come up with a modified alpha index [28], and we propose to call it the Omega index.

For a given numeric feature, i, the Omega index can be calculated as follows:

$$\Omega(i) = \frac{M_{icu} - M_{not\_in\_icu}}{\sigma_i} \tag{3}$$

Where,

M_{icu} is the median value of laboratory index, i, for COVID-19 infected patients who did require hospitalisation in ICU

$M_{not\_in\_icu}$ is the median value of laboratory index, i, for COVID-19 infected patients who did not require hospitalisation in ICU

σ_i is the standard deviation of laboratory index, i, for all COVID-19 infected patients

The numerator represents the difference in median values between patients who did not require ICU hospitalisation and those who did, normalising the impact of each laboratory index. The denominator scales the difference by the standard deviation of the laboratory index across

all patients. This scaling ensures that the index is not unduly influenced by extreme values and provides a standardised measure that accounts for the variability in the data distribution.

The selection of the median is based on its reduced sensitivity to extreme values or outliers when compared to the mean. When a dataset contains outliers, the mean can be significantly influenced, whereas the median remains less affected. Moreover, in distributions with skewed patterns, characterised by a prolonged tail on one side, the median often offers a more accurate representation of the central tendency.

The choice of using standard deviation is rooted in its ability to provide a more robust measure of variability compared to range due to its consideration of all values within the dataset.

The top 30 features ranked on the basis of the Omega index are depicted in Fig. 5.3. Similar to the alpha index, a positive omega index signifies that an increase in the laboratory index corresponds to ICU admission, while a negative index indicates that a decrease in the laboratory index predicts ICU admission.

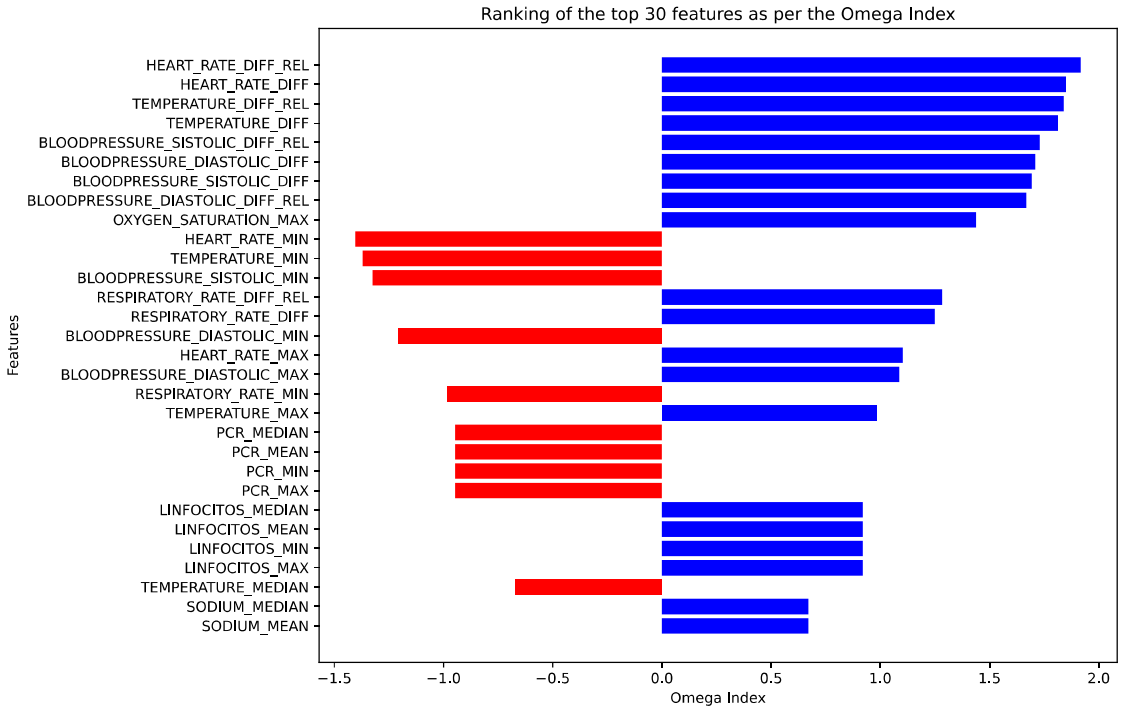

Fig. 5.3: Ranking of the top 30 features as per the Omega Index

5.3 Decision Trees

Decision Trees, a fundamental machine learning technique, have gained widespread recognition for their interpretability and simplicity, making them applicable across various domains, including healthcare. In the context of COVID-19 research, Decision Trees offer unique advantages in understanding and explaining the factors influencing ICU admission.

The core principle of Decision Trees involves recursively partitioning the data based on features to create a tree-like structure. Each node in the tree represents a decision or a split, and each leaf node provides the final prediction. This tree structure allows for straightforward interpretation, enabling healthcare professionals to easily grasp the decision-making process.

By examining the decision paths within the tree, healthcare professionals can gain a holistic understanding of the patient's condition and the key features driving the predictions. Decision Trees offer flexibility in handling various types of data, including categorical and numerical variables, without requiring extensive preprocessing. This versatility makes them particularly suitable for healthcare applications, where datasets often contain a mix of different data types.

The working of a decision tree is explained below:

1. **Selecting the Best Feature**: The algorithm starts with the entire dataset at the root node. It then evaluates different features and selects the one that best splits the data into distinct groups. The "best" feature is typically chosen based on criteria like information gain.
2. **Splitting the Dataset**: Once the best feature is chosen, the dataset is split into subsets based on the possible values of that feature. Each subset corresponds to a branch stemming from the root node.
3. **Recursive Partitioning**: The algorithm then repeats this process recursively for each subset created in the previous step. It continues splitting the data into smaller subsets until one of the stopping conditions is met, such as reaching a maximum depth, having a subset with only one class (in classification), or reaching a subset with a minimum number of samples.
4. **Creating Leaf Nodes**: Once the stopping conditions are met, the algorithm creates a leaf node at the end of each branch. This leaf node represents the final decision or prediction based on the majority class (in classification) or the mean value (in regression) of the samples in that subset.
5. **Handling Categorical and Numerical Features**: Decision trees can handle both categorical and numerical features. For categorical features, the algorithm can create

branches for each possible category. For numerical features, it selects thresholds to split the data into two subsets.

6. **Handling Overfitting**: Decision trees are prone to overfitting, especially when they have high depth or are not pruned. Techniques like limiting the maximum depth of the tree, setting a minimum number of samples required to split a node, or pruning the tree after it's built can help prevent overfitting.
7. **Making Predictions**: Once the decision tree is constructed, it can be used to make predictions on new, unseen data. The data traverses the tree from the root node to a leaf node based on the feature values, and the prediction is made based on the majority class or mean value of the samples in the leaf node.

The versatility of Decision Trees shines in capturing the heterogeneity observed in clinical presentations and factors influencing ICU admission for COVID-19 patients. Decision Trees can adapt to the intricate and varied nature of the disease, allowing for the identification of distinct decision paths for different patient profiles [29]. This adaptability ensures that the model can effectively handle the diversity of cases encountered in real-world healthcare settings, providing valuable insights into the nuanced relationships between features and outcomes.

They contribute to a human-centric approach in model development. Their simplicity and transparency make them accessible to healthcare professionals with varying levels of expertise. In the collaborative effort to combat COVID-19, Decision Trees serve as a bridge between machine learning insights and clinical understanding. This stands in contrast to black-box models, where the internal mechanisms are obscured, emphasising the value of human comprehension and involvement in the decision-making process for effective and trustworthy applications in healthcare [30].

CHAPTER 6: RESULTS AND ANALYSIS

Using an implementation of the decision tree algorithm in scikit-learn and clinical data from Hospital Sírio-Libanês, we developed a model aimed at enhancing ICU admission rates for COVID-19 patients. Subsequently, we conducted a comprehensive evaluation of the model's effectiveness and predictive capacity. Below, we present the results and provide detailed analysis of significant metrics to assess the precision and reliability of the model in predicting ICU admissions for COVID-19 patients.

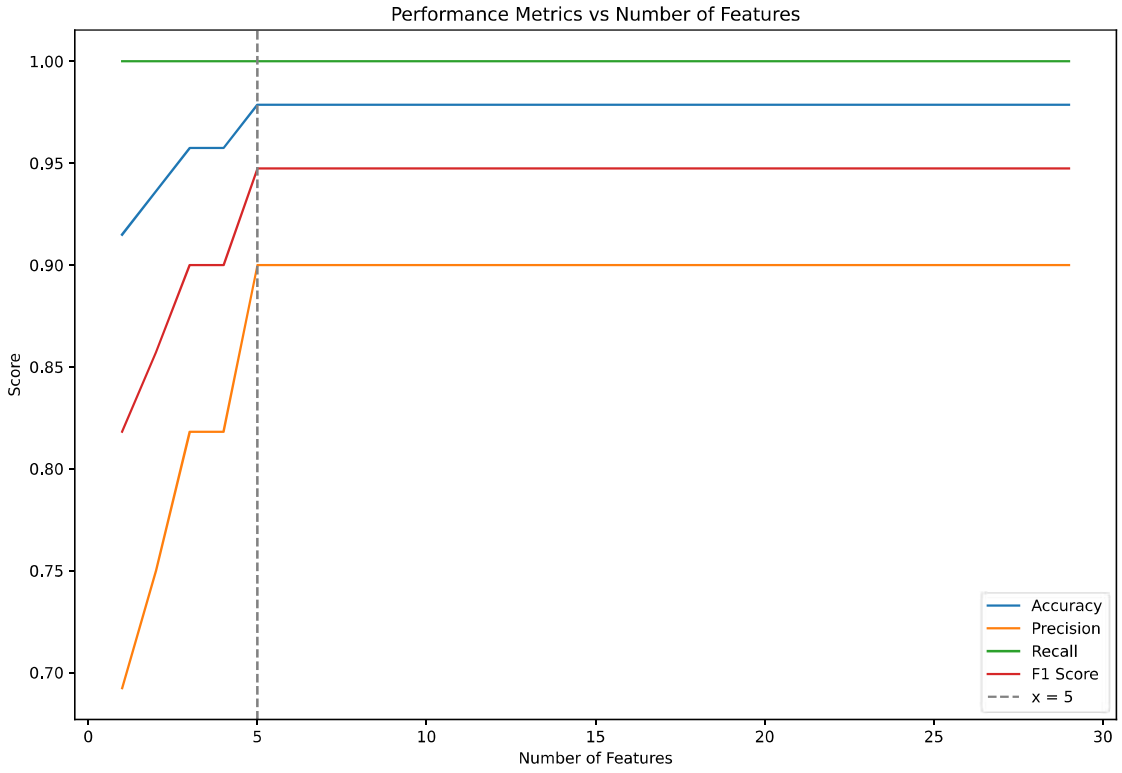

Fig. 6.1: Performance of the decision tree vs the number of features

From Fig. 6.1, we noted that the performance reaches saturation upon increasing the number of features beyond 5. Thus, we chose to proceed with 5 features as the ideal number of inputs to the Decision tree classifier. We further opted to use the decision criterion as 'entropy' as it provided us with more stable results.

TABLE 6.1: Results of the various performance metrics

Metric	Results
Accuracy	97.87 %
Precision	90.00 %
Recall	100.00 %
F1	94.74 %
Average Cross-Validation Accuracy (5-fold)	95.74 %
Mean Squared Error (MSE)	0.02
Root Mean Squared Error (RMSE)	0.15
R^2 (R-Squared)	0.86
ROC AUC	0.99

The decision tree model developed using scikit-learn and clinical data from Hospital Sírio-Libanês demonstrates promising performance metrics in predicting ICU admission rates for COVID-19 patients as depicted in Table 6.1. With an accuracy of 97.87%, the model achieves a high level of overall correctness in its predictions. Precision, which measures the proportion of true positive predictions among all positive predictions, stands at 90.00%, indicating a strong ability to correctly identify ICU admissions. The recall rate of 100.00% indicates that the model effectively captures all actual positive cases of ICU admissions. The F1 score, a harmonic mean of precision and recall, reaches 94.74%, suggesting a balanced performance between precision and recall.

Furthermore, the model's robustness is evidenced by an average cross-validation accuracy of 95.74% across five folds. Additionally, the mean squared error (MSE) and root mean squared error (RMSE) are low at 0.02 and 0.15 respectively, indicating minimal deviation between predicted and actual values. The model's R^2 (R-squared) value of 0.86 demonstrates a strong correlation between the predicted and observed outcomes. Lastly, the high ROC AUC score of 0.99 underscores the model's excellent ability to distinguish between positive and negative cases, further confirming its effectiveness in predicting ICU admissions for COVID-19 patients.

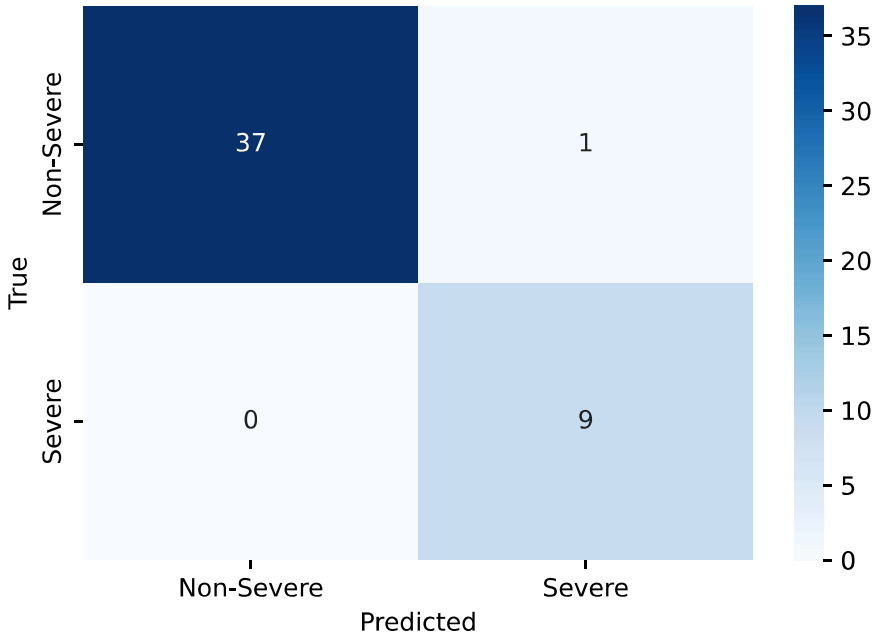

Fig. 6.2: Confusion matrix

In the confusion matrix depicted in Fig. 6.2, there are two classes: "Severe" and "Non-Severe". The diagonal cells show the number of correctly classified cases (37 + 9 = 46), while the off-diagonal cells show the number of cases that were misclassified (1 + 0 = 1).

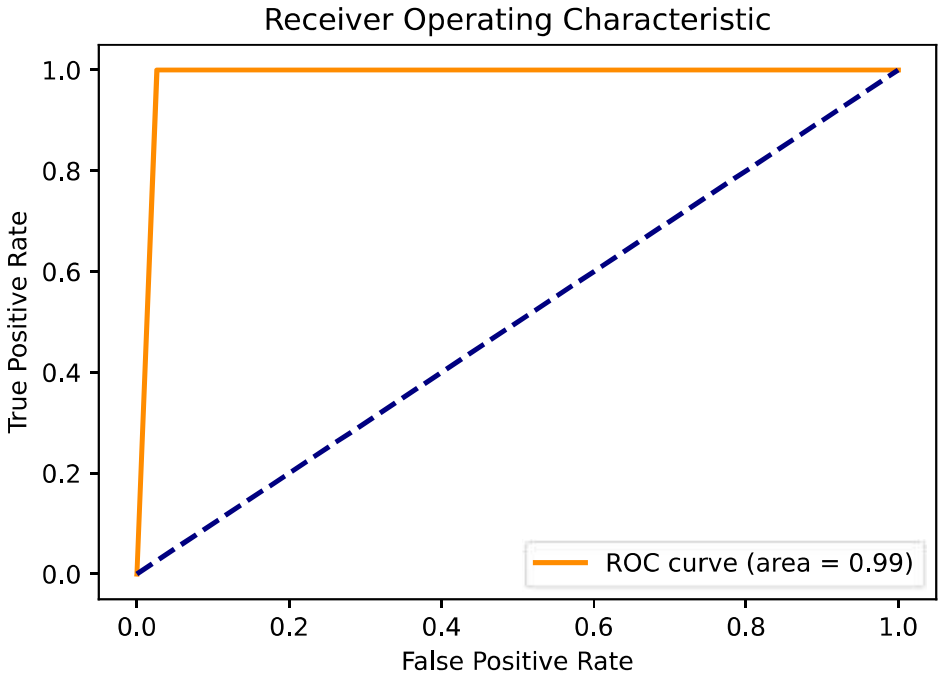

Fig. 6.3: Receiver Operating Characteristic

The ROC curve depicted in Fig. 6.3, has an area of 0.99, which implies that the model excels at distinguishing between positive and negative classes. Moreover, the convexity of the curve indicates that the TPR increases more rapidly than the FPR as the threshold is lowered, which is a good property for a ROC curve, as it means that the model can achieve a high TPR without too many false positives.

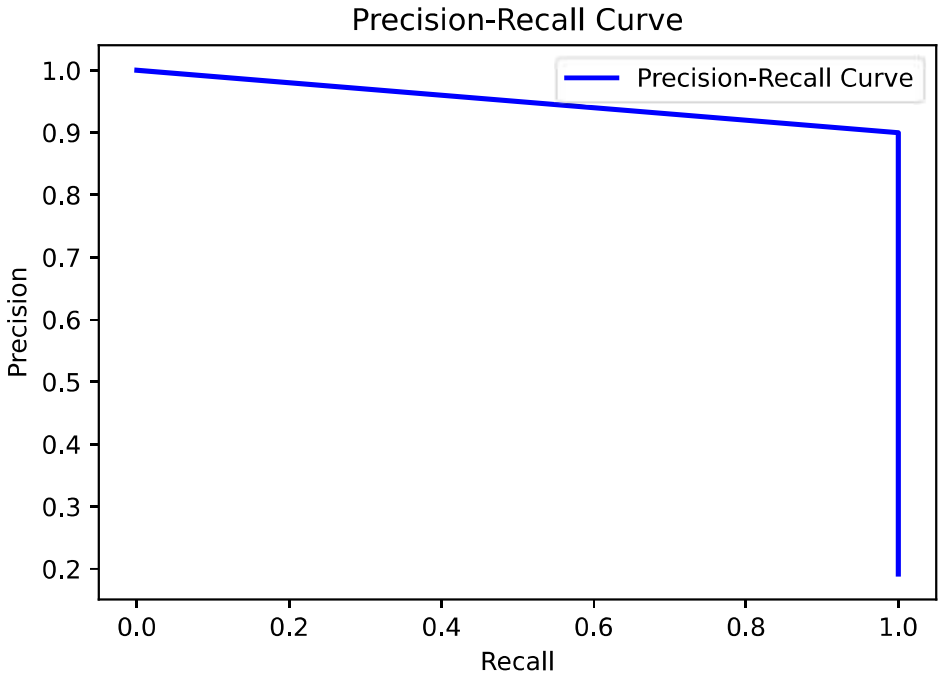

Fig. 6.4: Precision-Recall Curve

The precision curve depicted in Fig 6.4. shows that the model has a high precision for low recall values, which implies that the model is good at predicting positive cases when there are relatively few positive cases overall. However, the precision decreases as the recall increases, which means that the model is less accurate at predicting positive cases when there are more positive cases overall.

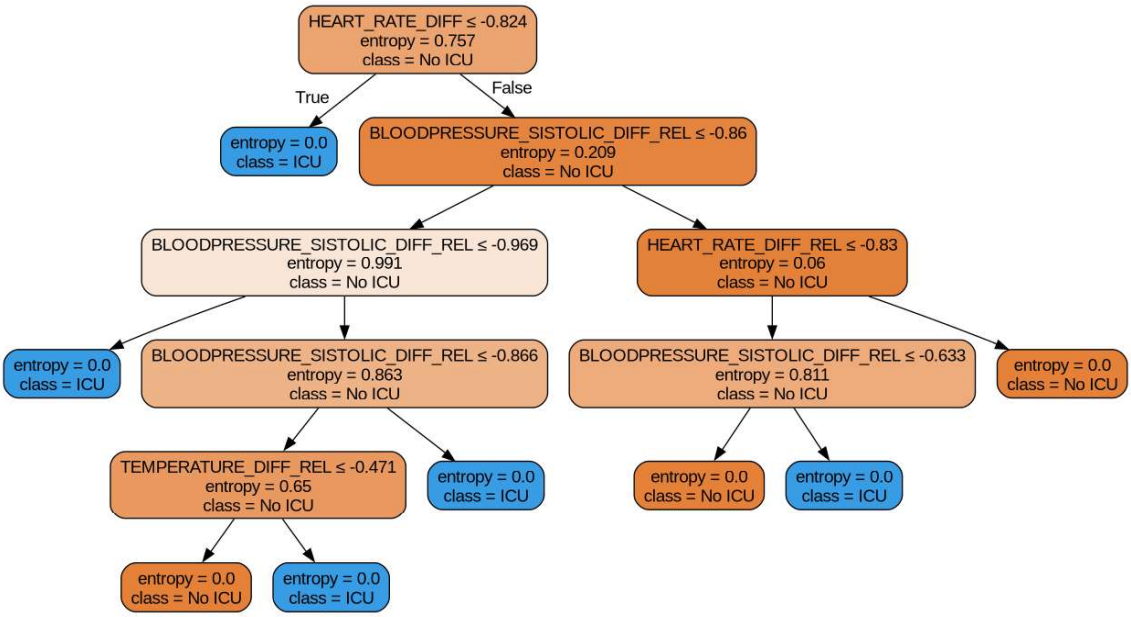

Fig. 6.5: The resulting Decision Tree

The decision tree depicted in Fig. 6.5 has been generated as a result of our analysis. Interpretability played a pivotal role in our approach, particularly in the medical domain where transparent insights are paramount for stakeholders. Decision Trees offer a hierarchical structure and clear decision logic, making them an ideal choice for understanding and communicating the model's reasoning effectively. This straightforward approach facilitates comprehension and ensures practicality, aligning well with the requirements of the dataset at hand.

CHAPTER 7: CONCLUSION AND FUTURE SCOPE

In conclusion, the decision tree model, implemented with clinical data from Hospital Sírio-Libanês using scikit-learn, demonstrates promising performance in predicting ICU admission rates for COVID-19 patients. Achieving an accuracy of 97.87% and precision and recall rates of 90.00% and 100.00% respectively, the model proves reliable in guiding medical practitioners in allocating ICU resources effectively. The interpretability of decision trees enhances trust in predictions and facilitates communication between stakeholders. Moving forward, refining the model with additional features and ensemble techniques could further enhance its performance and applicability across various healthcare settings, ensuring its continued efficacy in supporting medical decision-making processes amid evolving healthcare challenges, notably the ongoing COVID-19 pandemic.

7.1 Conclusion

- **Interpretability vs. Complexity Trade-off**: While decision trees offer interpretability, complex decision logic with numerous features can lead to convoluted trees that are challenging to interpret. In the context of the project, where transparency is crucial for stakeholders in the medical domain, excessively complex trees may hinder comprehension and limit trust in the model's predictions.

- **Limited Handling of Temporal Dynamics**: Decision trees inherently lack the ability to capture temporal dynamics or time-series patterns in the data. In the context of predicting ICU admission rates for COVID-19 patients, where disease progression and patient conditions may evolve over time, the model's inability to incorporate temporal information could result in suboptimal predictions.
- **Bias in Feature Importance**: Decision trees may exhibit bias in feature importance, favouring features that are inherently more predictive within the dataset. In the project, this bias could potentially overlook critical clinical indicators or covariates that are less prevalent but equally informative for predicting ICU admissions, leading to suboptimal model performance.
- **Limited Transferability to Different Settings**: Decision tree models trained on data from Hospital Sírio-Libanês may lack transferability to other healthcare settings or patient populations with different demographic characteristics, disease prevalence, or treatment

protocols. This limitation could hinder the model's generalizability and require extensive retraining or adaptation for deployment in diverse healthcare environments.

- **Ethical Considerations and Bias in Decision Rules**: The decision rules generated by decision trees may inadvertently reflect underlying biases present in the training data, potentially leading to discriminatory or inequitable predictions. In the project's context, where equitable allocation of ICU resources is paramount, it is crucial to assess and mitigate bias in the model's decision-making process to ensure fair and just outcomes for all patients.

7.2 Future Scope

- **Integration of Advanced Machine Learning Techniques**: While decision trees serve as a solid foundation, exploring more sophisticated machine learning techniques such as ensemble methods (e.g., Random Forests, Gradient Boosting) could enhance predictive performance and robustness. Ensemble techniques can mitigate the limitations of individual decision trees, offering improved accuracy and generalizability for predicting ICU admission rates among COVID-19 patients.
- **Incorporation of External Data Sources**: Leveraging additional external data sources, such as demographic data, geographic information, or epidemiological trends, could enrich the model's predictive capabilities. Integrating diverse datasets could provide a more comprehensive understanding of the factors influencing ICU admission rates and enable the model to adapt to evolving healthcare dynamics and regional variations.
- **Development of Explainable AI (XAI) Techniques**: Given the critical importance of transparency and interpretability in the medical domain, advancing Explainable AI (XAI) techniques tailored to decision tree models could further enhance stakeholders' understanding of the model's predictions. Techniques such as decision tree visualization, feature importance ranking, and rule extraction methods can provide insights into the model's decision-making process, fostering trust and facilitating informed decision-making by healthcare professionals.
- **Deployment of Real-time Predictive Analytics**: Moving towards real-time predictive analytics capabilities could empower healthcare providers with timely insights for proactive decision-making and resource allocation. By integrating the decision tree model into a scalable and responsive analytics platform, healthcare institutions can leverage up-to-date patient data to predict ICU admission rates in real-time, enabling more effective management of healthcare resources and patient care pathways.

- **Collaborative Research and Validation Studies**: Engaging in collaborative research initiatives and validation studies with multiple healthcare institutions and research partners could validate and refine the decision tree model across diverse clinical settings and patient populations. Conducting rigorous validation studies can enhance the model's generalizability, identify potential biases or limitations, and ensure its applicability in real-world clinical practice.
- **Ethical and Societal Impact Assessment**: Conducting comprehensive ethical and societal impact assessments of the decision tree model's deployment in clinical settings is essential to address potential biases, fairness concerns, and unintended consequences. Engaging with diverse stakeholders, including patients, healthcare providers, policymakers, and ethicists, can inform the development of responsible and equitable AI-driven healthcare solutions that prioritize patient well-being and societal values.

BIBLIOGRAPHY

[1] https://data.who.int/dashboards/covid19/cases

[2] https://data.who.int/dashboards/covid19/deaths

[3] Ceccato, A., Pérez-Arnal, R., Motos, A., Barbé, F. and Torres, A., 2022. One-year mortality after ICU admission due to COVID-19 infection. *Intensive Care Medicine*, *48*(3), pp.366-368.

[4] Aiyegbusi, O.L., Hughes, S.E., Turner, G., Rivera, S.C., McMullan, C., Chandan, J.S., Haroon, S., Price, G., Davies, E.H., Nirantharakumar, K. and Sapey, E., 2021. Symptoms, complications and management of long COVID: a review. *Journal of the Royal Society of Medicine*, *114*(9), pp.428-442.

[5] Shanbehzadeh, M., Kazemi-Arpanahi, H. and Nopour, R., 2021. Performance evaluation of selected decision tree algorithms for COVID-19 diagnosis using routine clinical data. *Medical Journal of the Islamic Republic of Iran*, *35*, p.29.

[6] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K. and Cilar, L., 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), p.e1379.

[7] Magazzino, C., Mele, M. and Coccia, M., 2022. A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality. *Epidemiology & Infection*, *150*, p.e168.

[8] S. Tiwari, P. Chanak and S. K. Singh, "A Review of the Machine Learning Algorithms for Covid-19 Case Analysis," in *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 1, pp. 44-59, Feb. 2023

[9] Bhosale YH, Patnaik KS. Application of Deep Learning Techniques in Diagnosis of Covid-19 (Coronavirus): A Systematic Review. Neural Process Lett. 2022 Sep 16:1-53.

[10] Bhosale, Y.H. and Patnaik, K.S., 2023. Bio-medical imaging (X-ray, CT, ultrasound, ECG), genome sequences applications of deep neural network and machine learning in diagnosis, detection, classification, and segmentation of COVID-19: a Meta-analysis & systematic review. *Multimedia Tools and Applications*, pp.1-54.

[11] Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C., Huang, L. and National Traditional Chinese Medicine Medical Team, 2022. Comparing different machine learning techniques for predicting COVID-19 severity. *Infectious diseases of poverty*, *11*(1), p.19.

[12] N. Rochmawati *et al*., "Covid Symptom Severity Using Decision Tree," *2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE)*, Surabaya, Indonesia, 2020, pp. 1-5,

[13] Kwekha-Rashid, A.S., Abduljabbar, H.N. and Alhayani, B., 2023. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. *Applied Nanoscience*, *13*(3), pp.2013-2025

[14] Bhatele, K.R., Jha, A., Tiwari, D., Bhatele, M., Sharma, S., Mithora, M.R. and Singhal, S., 2022. COVID-19 Detection: A Systematic Review of Machine and Deep Learning-Based Approaches Utilizing Chest X-Rays and CT scans. *Cognitive Computation*, pp.1-38.

[15] Heidari, A., Jafari Navimipour, N., Unal, M. and Toumaj, S., 2022. Machine learning applications for COVID-19 outbreak management. *Neural Computing and Applications*, *34*(18), pp.15313-15348.

[16] Gothai, E., Thamilselvan, R., Rajalaxmi, R.R., Sadana, R.M., Ragavi, A. and Sakthivel, R., 2023. Prediction of COVID-19 growth and trend using machine learning approach. *Materials Today: Proceedings*, *81*, pp.597-601.

[17] Das, S., Imtiaz, M.S., Neom, N.H., Siddique, N. and Wang, H., 2023. A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier. *Expert Systems with Applications*, *213*, p.118914.

[18] El Massari, H., Sabouri, Z., Mhammedi, S. and Gherabi, N., 2022. Diabetes prediction using machine learning algorithms and ontology. *Journal of ICT Standardization*, *10*(2), pp.319-337.

[19] https://ijcrt.org/papers/IJCRT22A6604.pdf

[20] Tazin, T., Alam, M.N., Dola, N.N., Bari, M.S., Bourouis, S. and Monirujjaman Khan, M., 2021. Stroke disease detection and prediction using robust learning approaches. *Journal of healthcare engineering*, *2021*.

[21] Shanbehzadeh, M., Nopour, R. and Kazemi-Arpanahi, H., 2022. Developing an artificial neural network for detecting COVID-19 disease. *Journal of Education and Health Promotion*, *11*.

[22] Elhag, A.A., Aloafi, T.A., Jawa, T.M., Sayed-Ahmed, N., Bayones, F.S. and Bouslimi, J., 2021. Artificial neural networks and statistical models for optimization studying COVID-19. *Results in Physics*, *25*, p.104274.

[23] Jain, G., Mittal, D., Thakur, D. and Mittal, M.K., 2020. A deep learning approach to detect Covid-19 coronavirus with X-Ray images. *Biocybernetics and biomedical engineering*, *40*(4), pp.1391-1405.

[24] Shanbehzadeh, M., Kazemi-Arpanahi, H. and Nopour, R., 2021. Performance evaluation of selected decision tree algorithms for COVID-19 diagnosis using routine clinical data. *Medical Journal of the Islamic Republic of Iran*, *35*, p.29.

[25] Yoo, S.H., Geng, H., Chiu, T.L., Yu, S.K., Cho, D.C., Heo, J., Choi, M.S., Choi, I.H., Cung Van, C., Nhung, N.V. and Min, B.J., 2020. Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging. *Frontiers in medicine*, *7*, p.427.

[26] Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z. and Kazemi-Arpanahi, H., 2022. Comparing machine learning algorithms for predicting COVID-19 mortality. *BMC medical informatics and decision making*, *22*(1), pp.1-12.

[27] https://www.kaggle.com/datasets/S%C3%ADrio-Libanes/covid19/data

[28] Asteris PG, Kokoris S, Gavriilaki E, Tsoukalas MZ, Houpas P, Paneta M, Koutzas A, Argyropoulos T, Alkayem NF, Armaghani DJ, Bardhan A, Cavaleri L, Cao M, Mansouri I, Mohammed AS, Samui P, Gerber G, Boumpas DT, Tsantes A, Terpos E, Dimopoulos MA. Early prediction of COVID-19 outcome using artificial intelligence techniques and only five laboratory indices. Clin Immunol. 2023 Jan;246:109218.

[29] Zhang, S., Zhang, K., Yu, Y., Tian, B., Cui, W. and Zhang, G., 2019. A new prediction model for assessing the clinical outcomes of ICU patients with community-acquired pneumonia: a decision tree analysis. *Annals of Medicine*, *51*(1), pp.41-50.

[30] Hakkoum, H., Abnane, I. and Idri, A., 2022. Interpretability in the medical field: A systematic mapping and review study. *Applied Soft Computing*, *117*, p.108391.