TUNISIAN REPUBLIC
Ministry of Higher Education and Scientific Research
University of Carthage

National Institute of Applied Sciences and Technology

Graduation Project

to obtain the

National Engineering Diploma
in Applied Sciences and Technology

Specialty: Industrial Computing and Automation

Early Prediction of Sepsis using Machine Learning Algorithms

Prepared by

Menyara Khaireddine

Hosted by Klinikum of Passau and University of Passau

Defended in front of a jury composed of:

Dr.	President of the jury
Dr.	Reviewer
Dr. Wiem Fekih Hassen	Supervisor at the University of Passau
Mr. Johannes Böhm	Supervisor at the Klinikum Passau
Prof. Imen Harbaoui	Supervisor at INSAT

School year: 2023 / 2024

TUNISIAN REPUBLIC
Ministry of Higher Education and Scientific Research
University of Carthage

National Institute of Applied Sciences and Technology

Graduation Project

to obtain the

National Engineering Diploma

in Applied Sciences and Technology

Specialty: Industrial Computing and Automation

Early Prediction of Sepsis using Machine Learning Algorithms

Prepared by

Menyara Khaireddine

Hosted by Klinikum of Passau and University of Passau

Supervisor at the University Dr. Wiem Fekih Hassen	Supervisor at the Clinic of Passau Mr.Johannes Böhm	Supervisor at INSAT Prof. Imen Harbaoui
Date :	Date :	Date :

School year : 2023 / 2024

Dedication

To my dearest father,
This journey has been one of immense growth and profound challenges, and through every step, your unwavering belief in my abilities has been a pillar of strength that I have continually relied upon. Your enduring presence through thick and thin, the sacrifices you've made, and the resilience you've shown have never gone unnoticed. Your enduring love and resilience have carried me through every challenge, every moment of doubt, and every achievement. I am profoundly grateful for instilling in me the strength and courage to become the strong woman I am today. I am immensely proud to call you my father, and I carry your lessons and love with me, always.

To my sister,
In you, I have found more than just a sister; you are my inspiration, my guide, and my idol. Your resilience and determination in the face of challenges have shaped my own aspirations. You have always been there, offering your support and wisdom, and your example has taught me the value of perseverance and grace under pressure. Thank you for lighting the way and for being the beacon I always look up to. Your influence in my life is immeasurable, and I am so proud to call you my sister.

***To my friends,**,*
To those from my hometown, who have stood by me since the very beginning, and to those I've met in my new town, who have welcomed me with open arms—thank you. The memories we've made, the challenges we've overcome together, and the unwavering support you've all provided have meant the world to me. You've been my family through every step, and this journey would not have been as fulfilling without each of you by my side.

Menyara Khaireddine

Dédicaces

À mon cher père,
Ce parcours a été une expérience de croissance immense et de défis profonds, et à chaque étape, ta foi inébranlable en mes capacités a été un pilier de force sur lequel je me suis constamment appuyée. Ta présence constante, les sacrifices que tu as faits et la résilience que tu as montrée n'ont jamais été ignorés. Ton amour et ta résilience ont porté à travers chaque défi, chaque moment de doute et chaque réussite. Je suis profondément reconnaissante de m'avoir inculqué la force et le courage de devenir la femme forte que je suis aujourd'hui. Je suis immensément fière de t'appeler mon père, et je porte toujours avec moi tes leçons et ton amour.

À ma sœur,
En toi, j'ai trouvé bien plus qu'une sœur; tu es mon inspiration et mon guide. Ta résilience et ta détermination face aux défis ont façonné mes propres aspirations. Tu as toujours été là, m'offrant ton soutien et ta sagesse, et ton exemple m'a appris la valeur de la persévérance. Merci d'avoir éclairé mon chemin, ton influence dans ma vie est inestimable, et je suis tellement fière de t'appeler ma sœur.

***À mes amis,**,*
À ceux qui m'ont soutenue depuis le tout début, et à ceux que j'ai rencontrés dans ma nouvelle ville, qui m'ont accueillie à bras ouverts—merci. Les souvenirs que nous avons créés, les épreuves que nous avons surmontées ensemble, et le soutien indéfectible que vous m'avez tous apporté ont été inestimables. Ce parcours n'aurait pas été aussi enrichissant sans chacun d'entre vous à mes côtés.

Menyara Khaireddine

Acknowledgment

I take immense pleasure in expressing my gratitude and deep appreciation to all who have contributed to the successful realization of this work. Your support, whether near or far, has been invaluable.

Firstly, I extend my sincere gratitude to ***Dr. Wiem Fekih Hassen*** *for her warm welcome and continuous support throughout my time at the University of Passau. Her invaluable advice and assistance have not only greatly enhanced the quality of this dissertation but also enriched my learning experience.*

I am also profoundly thankful to ***Mr. Johannes Böhm*** *at the Klinikum of Passau. His guidance and encouragement have been crucial in navigating the practical aspects of my research, providing a foundation for the clinical insights that have significantly shaped this work.*

Special thanks to ***Mrs. Imen Harbaoui****, my academic supervisor at INSAT, whose expertise and guidance have been instrumental in framing the academic rigor of my thesis. Her readiness to assist at every turn has greatly facilitated my research journey.*

I would also like to express my heartfelt appreciation to all the professors and academic staff at INSAT. Your dedication to nurturing students and fostering an environment of academic excellence has prepared me with the necessary knowledge and skills for my professional future.

Lastly, my deepest gratitude goes to the esteemed members of the jury, who have generously taken the time to assess this work. It is my sincere hope that this dissertation meets your expectations and reflects the high standards of quality and clarity you uphold.

Abstract

This thesis presents a machine learning approach for enhancing the early detection of sepsis, leveraging the predictive potential of routine blood test metrics and Procalcitonin (PCT) levels. The study employs advanced data analytics to examine the interrelations between blood count behaviors and PCT, identifying subtle patterns that signal the onset of sepsis. The developed model, demonstrated promising results with an accuracy of 86.52%, emphasizing its effectiveness in reliably predicting sepsis. This approach offers a cost-effective and dependable diagnostic tool, significantly reducing diagnosis time. It holds the potential to improve patient outcomes by enabling earlier and more precise interventions in sepsis treatment.

Keywords: Machine Learning, Sepsis Detection, Predictive Analytics, Blood Biomarkers, Hyperparameter Tuning, Blood Test Analysis, Healthcare Technology.

Contents

List of Figures

List of Tables

List of Acronyms

AdaBoost	=	**Ada**ptive **Boost**ing
AI	=	**A**rtificial **I**ntelligence
ANN	=	**A**rtificial **N**eural **N**etwork
AUROC	=	**A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic
BASOAB	=	**Baso**phils **Ab**solute Count
BASOPH	=	**Basoph**ils Percentage
CBC	=	**C**omplete **B**lood **C**ount
CRISP DM	=	**C**ross **I**ndustry **S**tandard **P**rocess **D**ata **M**ining
CRP	=	**C-R**eactive **P**rotein
EOSABS	=	**Eos**inophils **Abs**olute Count
EOSINO	=	**Eosino**phils Percentage
ERY	=	**Ery**throcytes Count
FN	=	**F**alse **N**egative
FP	=	**F**alse **P**ositive
GRANAB	=	**Gran**ocytes **Ab**solute Count
GRANUL	=	**Granul**ocytes Percentage
HB	=	**H**emoglobin
HK	=	**H**ematocrit
ICU	=	**I**ntensive **C**are **U**nit
IG-C	=	**I**mmature **G**ranulocytes **C**ount
IG-P	=	**I**mmature **G**ranulocytes **P**ercentage

• **IL6**	=	**I**nterleukin **6**
• **KNN**	=	**K**-**N**earest **N**eighbors
• **LEUKO**	=	**Leuko**cytes Count
• **LR**	=	**L**ogistic **R**egression
• **LSTM**	=	**L**ong **S**hort **T**erm **M**emory
• **LYMABS**	=	**Lympho**cytes **Abs**olute Count
• **LYMPHO**	=	**Lympho**cytes Percentage
• **MCH**	=	**M**ean **C**orpuscular **H**emoglobin
• **MCHC**	=	**M**ean **C**orpuscular **H**emoglobin **C**oncentration
• **MCV**	=	**M**ean **C**orpuscular **V**olume
• **ML**	=	**M**achine **L**earning
• **MODS**	=	**M**ulti **O**rgan **D**ysfunction **S**yndrome
• **MakroE**	=	**Macro**cytic **E**rythrocytes Percentage
• **MicroE**	=	**Micro**cytic **E**rythrocytes Percentage
• **MONABS**	=	**Mon**ocytes **Abs**olute Count
• **MONOZY**	=	**Mono**cytes Percentage
• **MPV**	=	**M**ean **P**latelet **V**olume
• **NB**	=	**N**aïve **B**ayes
• **PCA**	=	**P**rincipal **C**omponent **A**nalysis
• **PCT**	=	**P**rocalcitonin
• **PI**	=	**P**ermutation **I**mportance
• **PLT**	=	**P**latelets Count
• **qSOFA**	=	**q**uick **S**equential **O**rgan **F**ailure **A**ssessment
• **RBC**	=	**R**ed **B**lood **C**ells
• **RDW-SD**	=	**R**ed **C**ell **D**istribution **W**idth-**S**tandard **D**eviation
• **RF**	=	**R**andom **F**orest
• **RUSBoost**	=	**R**andom **U**nder **S**ampling **Boost**ing
• **SIRS**	=	**S**ystematic **I**nflammatory **R**esponse **S**yndrome

• **SOFA**	=	**S**equential **O**rgan **F**ailure **A**ssessment
• **SVM**	=	**S**upport **V**ector **M**achine
• **THROMB**	=	**Thromb**ocytes Count
• **TN**	=	**T**rue **N**egative
• **TR**	=	**T**rue **P**ositive
• **WBC**	=	**W**hite **B**lood **C**ells
• **XGBoost**	=	e**X**treme **G**radient **Boost**ing

General Introduction

Sepsis is a critical condition resulting from the body's extreme response to infection, often leading to organ failure [62]. It represents a global healthcare crisis, responsible for millions of deaths annually and presenting significant diagnostic challenges due to its rapid progression and nonspecific symptoms. Early detection is hampered by these factors, heightening the risk of severe complications and underscoring the need for accurate, rapid diagnostic methods.

Current diagnostic methods, based on clinical observation and biomarkers, are slow and costly, leading to delays that adversely affect patient outcomes. This highlights the urgent need for innovative tools that can swiftly and reliably predict sepsis onset, facilitating timely intervention.

Advancements in machine learning offer a promising solution by analyzing extensive clinical data to detect early signs of sepsis before symptoms are apparent. This thesis explores machine learning techniques using data from Klinikum Passau to develop predictive models that outpace traditional diagnostic methods in speed and accuracy.

The aim is to demonstrate how machine learning can significantly enhance patient outcomes in healthcare by accelerating sepsis diagnosis, thereby saving lives and easing healthcare burdens.

1

General Context

Introduction

In this chapter, we will provide a comprehensive overview of the foundational elements that underpin this research. We begin by introducing the host organizations which play critical roles in the academic and clinical aspects of this thesis. Following this, the chapter delves into the broader project context, highlighting the global healthcare challenge posed by sepsis and the complexities surrounding its early detection. We then outline the problem statement and project objectives. Finally, we present the management methodology, including a Gantt diagram that details the project timeline and key milestones.

1 Host Organization

In this section, we provide an overview of the key institutions that have contributed to the development of this thesis. These organizations, the University of Passau, DiMIS Group and Klinikum Passau, each play a vital role in their respective fields—academia and healthcare—offering a strong foundation of expertise and resources.

1.1 University Of Passau

The University of Passau [1], founded in 1978 and situated in the historic city of Passau, Germany, exemplifies a tradition of excellence in higher education. Recognized for its robust commitment to research, pedagogical innovation, and international collaboration, the university plays a pivotal role in shaping academic and professional futures .

The Figure 1.1 is the logo of the University of Passau. This emblem represents the university's identity and its dedication to academic excellence and innovation.

Figure 1.1: Logo of University of Passau

The University of Passau has a robust research ecosystem that significantly contributes to various academic fields, especially through its Faculty of Mathematics and Computer Science [2]. This faculty is the university's strongest in terms of securing external funding, with annual research grants amounting to approximately five million euros. Such funding supports a range of basic and applied research projects that not only advance academic knowledge but also have practical applications.

The faculty collaborates with both regional companies and global market leaders, enhancing the practical impact of its research. This connection between theoretical research and practical application ensures that the faculty's work remains relevant to current technological challenges and advancements.

[1] https://www.uni-passau.de/

[2] https://https://www.fim.uni-passau.de/

1.2 DiMIS Group

The Distributed Information and Multimedia Systems (DiMIS) group of the Faculty of Mathematics and Computer Science, led by Dr. Harald Kosch, focuses on research areas including metadata standardization, optimization of metadata coding, and the adaptation of multimedia streams. The department has played a significant role in developing an Moving Picture Experts Group (MPEG) query language. Their research also involves designing a platform for distributing interactive television and exploring its practical applications in household settings, particularly in conjunction with mobile devices. The logo of the hosting organization is shown in Figure 1.2.

Figure 1.2: Logo of DiMIS Group

1.3 Klinikum of Passau

Klinikum Passau [3], a key healthcare facility in Bavaria, Germany, provides comprehensive medical care with advanced technologies and a wide range of services. Known for its dedication to patient care and research, it serves as a crucial center for both emergency interventions and long-term health solutions. The hospital collaborates with various research institutions and universities, contributing to impactful medical studies and innovations. The dataset used in this thesis was collected from Klinikum Passau. The data comprises extensive patient records and treatment outcomes.

By integrating Klinikum Passau's dataset into this research, the thesis benefits from a robust foundation of clinically relevant data, facilitating a comprehensive analysis of treatment efficacies and patient outcomes. This collaboration with Klinikum Passau not only enriches the academic value of the research but also enhances the practical implications for future medical practices.

[3] https://www.klinikum-passau.de/

The Figure 1.3 is the logo of Klinikum Passau, which symbolizes the hospital's commitment to excellence in healthcare and its central role in medical innovation and patient care in Bavaria.

Figure 1.3: Logo of Klinikum of Passau

2 Project Presentation

In this section, we delve into the landscape of healthcare challenges, where sepsis stands out due to its severe impact on patient mortality and systemic healthcare burdens globally. Sepsis, a critical and often fatal response to infection leading to organ dysfunction, poses a significant global healthcare challenge due to its rapid escalation from initial symptoms to severe stages and its profound impact on patient mortality and healthcare burdens [28]. The early detection of sepsis is complex due to the variability of its symptoms and the lack of a definitive, singular diagnostic test. Traditional methods, which rely heavily on specific biomarkers and comprehensive clinical evaluations, are not only time-consuming but also costly. These limitations often result in delayed diagnoses, adversely affecting patient outcomes and increasing healthcare costs. Given the high stakes of timely and accurate sepsis detection, there is a pressing need for more efficient diagnostic solutions [19]. Advances in ML offer a promising avenue for enhancing the diagnostic capabilities of healthcare systems.

2.1 Project Context

2.1.1 Sepsis Statistics and Economic Impact

Sepsis remains a formidable challenge in global healthcare, with its high incidence and mortality rates significantly contributing to health burdens worldwide. As of 2020, it was estimated that sepsis accounted for approximately 48.9 million cases globally each year, resulting in around 11 million deaths, representing nearly 20% of all global deaths annually [44]. A systematic review and meta-analysis by Fleischmann et al. (2016) corroborates these figures, reporting a global death rate of 19.4% [23]. This mortality rate varies significantly by region, with high-income countries experiencing a rate of around 15.7%, while low- and middle-income countries face a staggering 34.7%. Angus et al. (2001) found that the mortality rate for severe sepsis and septic shock ranges between 40% and 60% [5].

2.1.2 Long-Term Consequences of Sepsis

In hospital settings, the impact of sepsis is particularly acute; it is estimated that for every 1000 patients hospitalized, about 15 develop sepsis as a direct complication of their healthcare treatment. Survivors of sepsis often endure long-term physical, cognitive, and psychological impairments, which contribute to a substantial burden of morbidity. The Surviving Sepsis Campaign (SSC) found that one year after discharge, 33% of over 1000 sepsis survivors experienced cognitive dysfunction, 43% had new functional limitations, and 27% exhibited symptoms of Post-Traumatic Stress Disorder (PTSD) [11]. These long-term complications significantly affect survivors' quality of life and result in considerable healthcare needs.

2.1.3 Burden on Healthcare Systems

Sepsis also places a substantial burden on healthcare systems worldwide, a challenge intensified by its high hospitalization rates, extended hospital stays, and significant resource utilization. Treating sepsis often requires intensive medical intervention, including prolonged Intensive Care Unit (ICU) stays, advanced medical technologies, and a multidisciplinary approach, which drives up the overall cost of healthcare services. According to a World Health Organization (WHO) study, the economic impact is particularly severe in high-income countries, where the average cost per sepsis patient exceeds USD 32,000 [68]. A study made in 2018 estimated that the total annual cost of sepsis hospitalizations in the United States was approximately $24 billion, accounting for 13.3% of all hospital expenditures [43].

Furthermore, sepsis is associated with a higher risk of readmissions and hospital-acquired infections, adding additional pressure to healthcare systems. Rhee et al. (2017) found in a retrospective cohort study that sepsis survivors had a 38% higher risk of hospital readmission within 90 days compared to non-sepsis patients [41].

2.2 Problem Statement

The primary challenge this research addresses is developing an efficient, reliable, and cost-effective method for early sepsis diagnosis using ML technologies. Traditional diagnostic methods, including the use of PCT as a biomarker, are highly effective but often hindered by their high costs and the time it takes to obtain results. These methods are typically employed only when there is a suspicion of sepsis, which can lead to potential delays in diagnosis [70]. The delays are problematic as sepsis requires rapid intervention to prevent severe complications and increase survival rates. Consequently, there is a significant need for an innovative approach that enhances the timeliness and accuracy of sepsis detection.

2.3 Proposed Solution

This study aims to leverage the ubiquity and rapid availability of regular blood tests performed in hospitals—tests that are less expensive and routinely administered to all hospitalized patients. By focusing on the interrelation between blood count behavior and PCT levels—a reliable biomarker for sepsis—our approach seeks to significantly refine the accuracy and speed of sepsis diagnostics. This method analyzes routine blood tests and PCT measurements conducted in hospital settings. These tests, which are less expensive and frequently administered, provide a rich data source for early detection.

The core of our proposed solution is a Machine Learning (ML) classification model, where algorithms analyze blood count data to identify subtle patterns that indicate the onset of sepsis. This analysis categorizes patients into different risk categories, effectively predicting the likelihood of developing sepsis. This strategy aims at a revolutionary shift in how this critical condition is diagnosed and managed. The Figure 1.4 will clearly illustrate the end-to-end process of data input through to clinical decision-making.

The workflow graph for the proposed ML model would start with inputs of routine blood tests and demographic data on one side. These inputs feed into the ML algorithm module, which processes the data to detect patterns and correlations associated with sepsis risk. The output from this module would be a quantified risk percentage, categorizing patients according to their likelihood of developing sepsis. This quantified risk serves as a crucial guide for medical professionals in making informed decisions about further diagnostic or therapeutic steps.

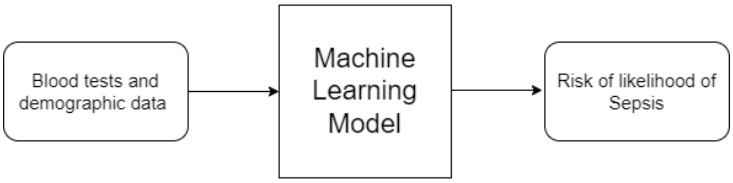

Blood tests and demographic data
Machine Learning Model
Risk of likelihood of Sepsis

Figure 1.4: Machine Learning Model Workflow

2.4 Project Objectives

The main objectives of the project are outlined as follows:

1. **Analyze Blood Features:** Understand and quantify the significance of each feature in the blood dataset, particularly its influence on PCT levels and sepsis diagnosis. Focus on the interrelation between blood count behavior and PCT levels.
2. **Dataset Extraction:** Identify and extract several relevant sub-datasets tailored for targeted analysis and model training.
3. **Algorithm Testing:** Apply different ML algorithms to these datasets to ascertain the most effective approaches for sepsis prediction.
4. **Hyperparameter Tuning:** Perform hyperparameter tuning on the most promising models, considering their accuracy, processing performance, and rates of false positives and false negatives.
5. **Benchmarking:** Compare the results of our ML models to existing baseline studies with similar objectives to gauge advancements and improvements.
6. **Evaluate Predictive Power:** Assess the efficacy of using routine blood test features as predictors for sepsis, aiming to validate or redefine their utility in clinical settings.

3 Management Methodology

3.1 Overview of CRISP-DM Methodology

The Cross Industry Standard Process for Data Mining framework (CRISP-DM) [1] provides a robust and systematic approach tailored for data mining projects like our study.

This iterative methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The Figure 1.5 shows each phase iteratively refining its predecessors through continuous learning and validation aligned with the project's business objectives.

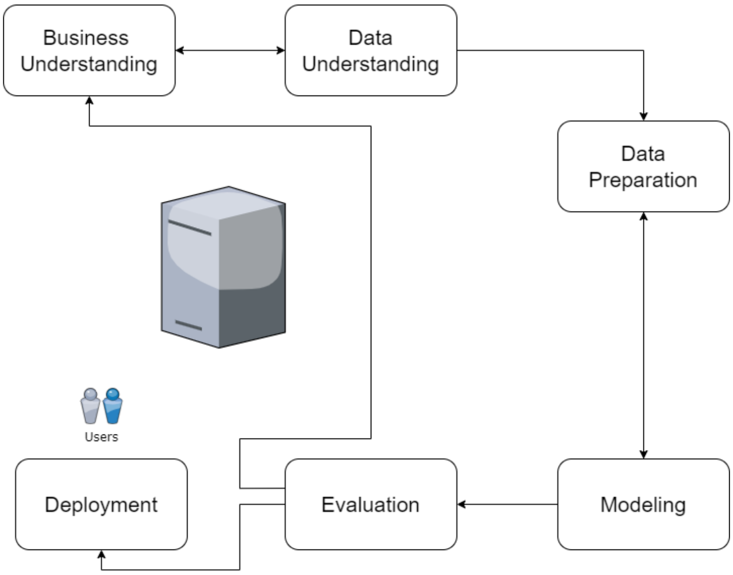


Figure 1.5: Workflow of CRISP-DM Method

This cyclical refinement is especially crucial in our dynamic, data-driven endeavor where early detection of sepsis is the focal challenge. CRISP-DM's adaptability allows for responsive adjustments as insights evolve, ensuring that the model remains aligned with the pressing needs of healthcare settings and improves outcomes in sepsis management.

3.2 Rationale for Choosing CRISP-DM

Choosing the appropriate methodology for a project that intersects healthcare and data science is pivotal. The CRISP-DM methodology is preferred over traditional models like the Waterfall methodology for several reasons. CRISP-DM's iterative nature supports the evolving requirements of data-driven healthcare projects by allowing for continuous refinement and learning. This is particularly beneficial in a field like sepsis detection where new data can alter initial understandings and necessitate quick adjustments to the modeling approach. Below is the Table 1.1 comparing CRISP-DM with a traditional methodology : Waterfall, to highlight why CRISP-DM is more suited for our project on early sepsis detection using machine learning technologies.

Table 1.1: Comparison of CRISP-DM and Waterfall Methodologies

Feature	**CRISP-DM**	**Waterfall (Traditional Approach)**
Flexibility	High, allows iterative refinements based on insights.	Low, sequential with no overlap between phases.
Adaptability to Change	Highly adaptable, can incorporate changes at any phase.	Poor, changes are difficult once a phase is completed.
Feedback Integration	Continuous throughout the project lifecycle.	Generally only after the completion of the project.
Project Management	Cyclical process that revisits and refines previous steps.	Linear approach, follows strict order without revisiting.
Suitability for Complex Projects	Ideal for complex, unpredictable projects like sepsis detection.	Best for projects with well-defined, unchanging requirements.

3.3 Project Timeline: Gantt Chart Representation

To effectively manage and visualize the timeline of our sepsis prediction project, the Gantt chart of Figure 1.6 has been developed. This chart provides a comprehensive overview of the project's phases in alignment with the CRISP-DM methodology, illustrating the start and end dates for each phase, including overlaps between tasks.

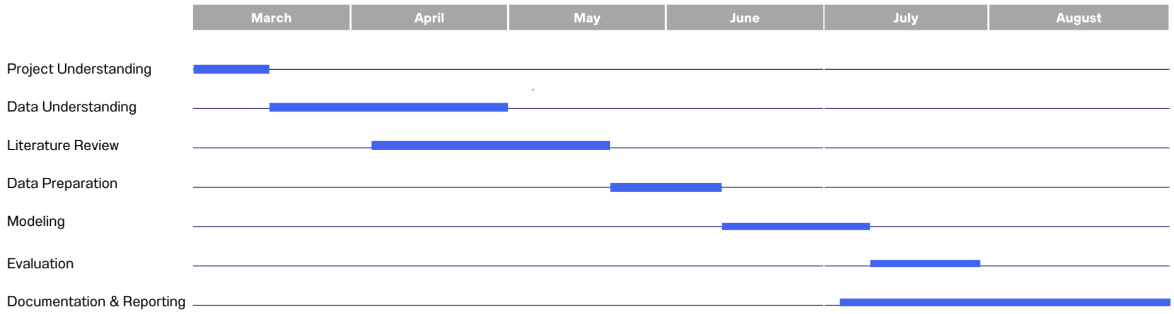


Figure 1.6: Gantt Chart of Sepsis Predicition Project

This is the detailed breakdown of each task outlined in the project's timeline:

- Project and Data Understanding: Define and clarify the project scope and objectives. Collect initial datasets from collaborating healthcare facilities, perform exploratory data analysis to assess the data quality, and identify preliminary patterns or anomalies.
- Literature review: Conduct meetings with healthcare professionals to gather insights on the current challenges in sepsis diagnosis, and review literature and papers to align the project with the latest research findings and healthcare standards.

- Data Preparation: Clean the data by addressing missing values and outliers, integrate various data sources, and transform data into suitable formats for machine learning models.
- Modeling: Conduct feature selection, select appropriate machine learning algorithms and train models. This phase also involves optimizing the models through hyperparameter tuning to enhance their predictive accuracy.
- Evaluation and Benchmarking: Evaluate model effectiveness, focusing on sensitivity, specificity, and the overall accuracy. Adjust models based on evaluation results and compare outcomes with existing sepsis prediction methodologies.
- Documentation and Reporting: Compile detailed documentation of all phases, including methodologies used, model performance data, and insights gained during the project. Prepare the final project report that encapsulates the research, findings, and practical implications for sepsis detection.

Conclusion

This chapter has provided a detailed context for the research, establishing the significance of the host organizations, the critical nature of the sepsis challenge, and the innovative approaches being explored to address it. By laying out the project objectives and management methodology, including the Gantt diagram, we have set the stage for the subsequent chapters that will delve into the technical and analytical aspects of this work.

2

Background

Introduction

In this chapter, we provide a comprehensive overview of sepsis, emphasizing its clinical significance and the challenges associated with its diagnosis. We begin by defining sepsis, exploring its causes, and tracing the evolution of its diagnostic criteria over time. Additionally, we discuss the widespread impact of sepsis on global healthcare systems, highlighting its significant burden in terms of mortality, morbidity, and resource utilization. This chapter also includes a literature review, offering a state-of-the-art examination of the latest scientific advancements in sepsis diagnosis and management.

1 Sepsis: Clinical Significance and Diagnostic Challenges

In this section, we examine the clinical significance and diagnostic challenges associated with sepsis. The section begins by defining sepsis and exploring its causes, tracing the evolution of its definition and diagnostic criteria over time. We then discuss the progression of sepsis, detailing how the condition advances from initial infection to severe organ dysfunction and, ultimately, septic shock.

1.1 Definition of Sepsis

Sepsis is a life-threatening condition that arises when the body's response to an infection causes widespread inflammation, leading to tissue damage and organ failure. The initial Systematic Inflammatory Response Syndrome (SIRS), is often triggered by an initially localized infection such as pneumonia, urinary tract infections, or infected wounds. The body's immune system, attempting to fight the infection, might go on overdrive, which may result in harm far beyond the original infection site [62].

The consensus definitions have evolved significantly over the years. Initially defined in 1991 (Sepsis-1) by the American College of Chest Physicians (ACCP) and the Society of Critical Care Medicine (SCCM), sepsis was considered a systemic inflammatory response to infection [8]. This definition was refined in 2001 (Sepsis-2), maintaining the emphasis on the inflammatory response but with clearer criteria for organ dysfunction [33]. The most transformative change came with Sepsis-3 in 2016, which shifted focus from the inflammatory response to organ dysfunction and a dysregulated host response to infection. This latest definition describes sepsis as life-threatening organ dysfunction caused by a dysregulated response to infection, quantified by an increase in the Sequential Organ Failure Assessment (SOFA) score of two points or more [63]. This evolution of the definition of sepsis reflects growing insights into the pathophysiology of sepsis and improvements in diagnostic criteria.

1.2 Causes of Sepsis

Sepsis is most commonly triggered by bacterial infections, but it can also be caused by viral or fungal infections. The pathogens enter the body through various routes; common examples include pneumonia, urinary tract infections, abdominal infections, and primary bloodstream infections.

- **Bacterial Infections:** Bacterial infections are the most frequent cause of sepsis. These bacteria can invade the body from an external source, or they can be part of the body's normal flora that become harmful under certain conditions, such as reduced immunity [20].
- **Viral Infections:** Viral infections leading to sepsis are increasingly recognized, especially in the cases of influenza and COVID-19. These viruses precipitate severe inflammatory responses in the body, leading to septic conditions particularly in individuals with compromised immune systems or existing health conditions [6].
- **Fungal Infections:** Fungal infections such as those caused by *Candida* species can also lead to sepsis, especially in immunocompromised patients, including those undergoing chemotherapy, taking steroids, or with invasive devices like catheters. These infections are particularly dangerous as they are more resistant to treatment and can rapidly progress to severe sepsis [13].
- **Other Precipitating Factors:** In addition to infections, several non-infectious factors can precipitate the inflammatory cascade that leads to sepsis. These include but are not limited to severe burns, significant trauma, and post-surgical complications. The presence of foreign bodies, such as catheters and surgical implants, can also serve as a focal point for infection, thus increasing the risk of developing sepsis [22].

1.3 Progression and Symptoms of Sepsis

In understanding the multifaceted nature of sepsis, it is crucial to acknowledge its dynamic progression through several distinct stages, as elucidated by Dobson and colleagues [17]. Their work delineates the trajectory from the initial microbial entry to the severe endpoint of MODS. The study outlines how an initial infection escalates from a systemic inflammatory response (SIRS) to a complex immune dysregulation, eventually leading to widespread organ failure. Building on this foundation, let us explore each stage of sepsis [14] to gain a deeper understanding of the processes that underpin its development:

1. **Entry of Microbes:** Infection begins when harmful microbes enter the body through various routes such as wounds or inhalation.
2. **SIRS:** The body's initial response involves a widespread inflammatory state intended to contain and combat the infection.
3. **Dysregulation of Immune Response:** If the infection is not adequately controlled, the immune response becomes dysregulated, leading to excessive cytokine production and nitric oxide release. This results in vasodilation and increased capillary permeability, causing a drop in blood pressure and fluid leakage into tissues.
4. **Coagulation Cascade Activation:** The inflammatory response concurrently activates the coagulation cascade, resulting in the formation of microclots that obstruct blood flow and exacerbate tissue hypoxia.
5. **Reduced Organ Perfusion and Function:** Diminished blood flow leads to reduced oxygen and nutrient delivery to organs, impairing their function. The heart attempts to compensate by increasing output; however, this can lead to cardiac failure.
6. **MODS:** Persistent hypoperfusion and a hyperinflammatory state culminate in multiple organ failures, known as MODS.

The accompanying Figure 2.1 adapted [61] outlines the sequential progression of sepsis and its escalation.

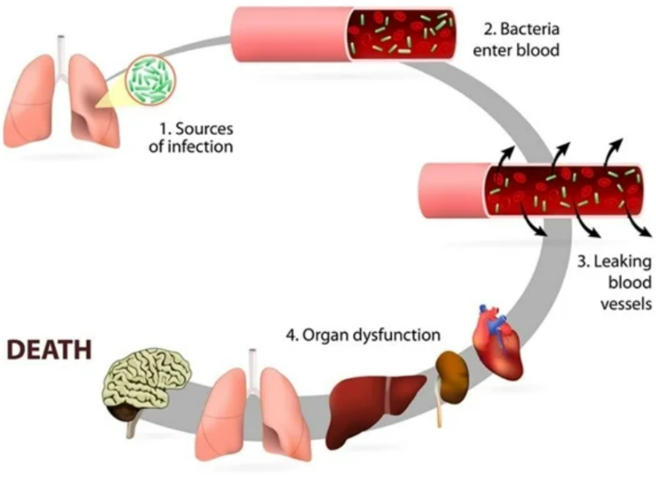


Figure 2.1: Stages of Sepsis

The progression of sepsis is systematically categorized into distinct stages, each indicating an escalation in severity. This stratification of sepsis stages is detailed extensively by Siddharth Dugar [18], who outline the clinical criteria and management strategies pertinent to each stage, from SIRS through to septic shock.

Sepsis can be categorized into 4 distinct stages [18], each reflecting an escalation in severity:

1. **SIRS:** SIRS begins when microbes enter the body, triggering an immune response. The initial signs include general symptoms like fever, chills, and a rapid heart rate, reflecting the body's widespread inflammatory state intended to combat the infection. This stage is directly linked to the initial immune response, where the body attempts to contain the infection through a generalized inflammatory reaction.
2. **Sepsis:** If the SIRS response fails to control the infection, the immune system becomes dysregulated, leading to excessive inflammation. The condition escalates to sepsis when SIRS is accompanied by signs of organ dysfunction, such as hypotension or altered organ perfusion, which are outcomes of the immune dysregulation and subsequent inflammatory damage.
3. **Severe Sepsis:** The activation of the coagulation cascade and the resultant formation of microclots obstruct blood flow, exacerbating tissue hypoxia and leading to reduced organ perfusion. Progression to severe sepsis is marked by acute organ dysfunction or hypoperfusion abnormalities. Signs include hypotension, high lactate levels, or decreased urine output due to impaired kidney function.
4. **Septic Shock:** The culmination of persistent hypoperfusion and a hyperinflammatory state leads to multiple organ failures. Septic shock is the most severe stage, characterized by severe hypotension despite fluid resuscitation, requiring vasopressors to maintain blood pressure. Other symptoms include flushed and cold skin, mottling, bluish discoloration, and the presence of multiple organ dysfunction.

The accompanying Figure 2.2 adapted [65] succinctly illustrates the causes and progression of sepsis from initial infection through to the severe stages of septic shock. It highlights the transition from SIRS, through the stages of sepsis and severe sepsis, to the critical point of septic shock, each marked by increasing severity and risk of death.

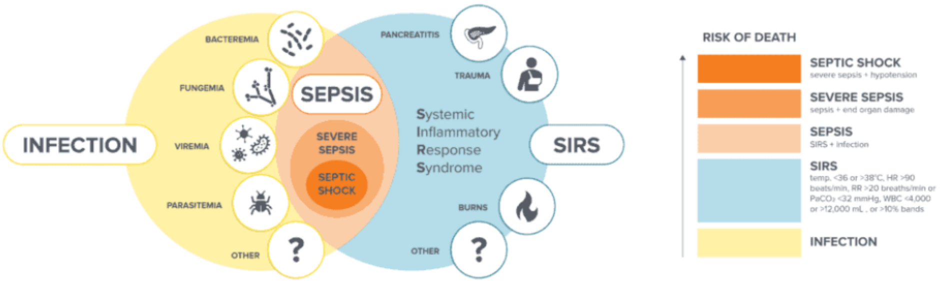

BACTEREMIA
FUNGEMIA
INFECTION
VIREMIA
PARASITEMIA
OTHER
?
SEPSIS
SEVERE SEPSIS
SEPTIC SHOCK
PANCREATITIS
TRAUMA
Systemic Inflammatory Response Syndrome
SIRS
BURNS
OTHER
?
RISK OF DEATH
SEPTIC SHOCK
severe sepsis + hypotension
SEVERE SEPSIS
sepsis + end organ damage
SEPSIS
SIRS + infection
SIRS
temp. <36 or >38°C, HR >90 beats/min, RR >20 breaths/min or PaCO₂ <32 mmHg, WBC <4,000 or >12,000 mL , or >10% bands
INFECTION

Figure 2.2: Progression of Sepsis

2 Conventional Approaches to Sepsis Diagnostic

2.1 Clinical Evaluation

The clinical evaluation is the foundational step in diagnosing sepsis, involving a thorough assessment by healthcare providers to identify signs of infection and symptoms indicative of sepsis, such as fever, increased heart rate, increased respiratory rate, confusion or cognitive changes, and hypotension [42]. During this evaluation, physicians conduct a physical examination to inspect the skin for signs of infection, palpate for abdominal tenderness, and listen to the lungs and heart to assess respiratory and cardiac function. Risk factors such as age, immunocompromised states, chronic diseases, and recent surgeries are also considered. In some settings, structured screening tools may be used. This initial clinical assessment is critical, guiding subsequent diagnostic tests and treatment strategies, setting the stage for timely interventions that can significantly improve patient outcomes.

2.2 Laboratory Tests

Several laboratory tests are pivotal in the diagnosis and management of sepsis, helping to identify the presence of an infection, evaluate organ function, and guide treatment decisions [24]:

- **Blood Cultures:** These are crucial for detecting the presence of bacteria or fungi in the bloodstream, serving as direct indicators of systemic infection [36]. Blood cultures help identify the specific pathogens responsible for sepsis, which is essential for selecting the most effective antibiotic treatment .

- **Lactate Levels:** High levels of lactate may indicate that body tissues are not receiving enough oxygen, a common issue in sepsis [15].
- **C-Reactive Protein (CRP):** A protein produced by the liver in response to inflammation, elevated CRP levels are used to detect inflammation and monitor treatment response [45].
- **Complete Blood Count (CBC):** This test includes various components such as Red Blood Cells (RBC), White Blood Cells (WBC), Hemoglobin (HB), Hematocrit (HK), Platelets (PLT) [4].
- **PCT:** PCT is a biomarker that rises in response to bacterial infections. Among various markers used for diagnosing bacterial infections, PCT is considered the most reliable to date [71]. Its specificity in indicating bacterial infection makes it an essential tool in clinical settings, particularly for guiding antibiotic therapy decisions .

2.3 Imaging Tests

Imaging tests may be employed to locate the source of infection, especially when the initial infection site is not apparent from physical examination or initial laboratory tests. X-rays are frequently used to detect respiratory infections like pneumonia by revealing lung abnormalities [16]; Computed Tomography Scans (CT Scans) offer detailed images, essential for identifying intra-abdominal infections such as appendicitis or abscesses. Magnetic Resonance Imaging (MRI) provides excellent contrast resolution for diagnosing soft tissue infections and is particularly useful for complex cases involving the spine or brain. Ultrasounds serve well in evaluating soft tissue structures and fluid status, ideal for detecting infections in organs like the gallbladder or kidneys and are invaluable in critical care for assessing cardiac function.

2.4 Sepsis Scores and Criteria

To assess the severity of sepsis and the risk of progression to septic shock, tools like the Sequential Organ Failure Assessment (SOFA) and quick Sequential Organ Failure Assessment (qSOFA) scores are utilized. These tools evaluate the extent of organ dysfunction and are valuable in both predicting outcomes and guiding treatment interventions.

SOFA Score: Measures organ function across six systems (respiratory, coagulation, liver, cardiovascular, central nervous system, and renal) with scores ranging from 0 (normal function) to 4 (high degree of dysfunction) [3]. A rise in the SOFA score by two or more points is associated with increased mortality risk.

qSOFA Score: A simplified tool designed for quick assessment outside the ICU, focusing on three criteria—respiratory rate, altered mentation, and systolic blood pressure—to identify patients at greater risk of sepsis [3].

3 Diagnostic Challenges and Early Detection

3.1 Challenges in Early Detection

- **Variability of Symptoms:** Sepsis presents with a wide array of possible symptoms that can vary greatly between patients [10], depending on factors such as the underlying cause of the infection, the patient's age, immune status, and presence of comorbid conditions. Common symptoms include fever, chills, rapid breathing, confusion, and cardiovascular instability; however, these are not exclusive to sepsis and can be seen in many other medical conditions. This symptom overlap often leads to diagnostic confusion, especially in early stages when symptoms are mild or nonspecific. Moreover, sepsis can progress rapidly, with patients deteriorating quickly from a mild, seemingly manageable state to severe sepsis or septic shock, necessitating urgent and accurate diagnosis.

- **Lack of a Singular Diagnostic Test:** Currently, there is no single laboratory test that can definitively diagnose sepsis, adding another layer of complexity to its early detection. While tests like blood cultures, lactate levels can support a diagnosis of sepsis, they are not foolproof and often require interpretation in the context of the patient's overall clinical picture. Blood cultures, for instance, which are the gold standard for identifying the causative organism, have variable sensitivity and can take time to yield results. High lactate levels can indicate tissue hypoperfusion associated with sepsis, but elevated lactate can also result from other conditions not related to infection. Notably, while PCT is the most reliable marker available for diagnosing bacterial infections related to sepsis, its use is constrained by cost considerations and time, as results are not immediate. This can be problematic given that a timely diagnosis is critical for effective sepsis management [19].

3.2 Consequences of Delayed Diagnosis

- **Increased Morbidity:** Delayed diagnosis of sepsis can lead to a rapid progression of the disease, resulting in the development of severe sepsis and septic shock. This progression often leads to MODS, where multiple organ systems begin to fail. The longer the delay in diagnosis and treatment, the greater the risk and extent of organ damage [38]. This can result in prolonged hospital stays, increased need for intensive care, and long-term disabilities. Conditions such as acute renal failure, respiratory failure, and severe sepsis-induced cardiomyopathy are examples of complications that may require prolonged medical intervention, including dialysis, mechanical ventilation, and other supportive measures which significantly impact patient quality of life.

- **Increased Mortality:** Mortality rates for sepsis increase with each hour that treatment is delayed. Studies have shown that the mortality rate for septic shock can increase by as much as 7.6% [32] for each hour that the administration of antibiotics is delayed after the onset of hypotension. Early intervention is crucial; for instance, the administration of appropriate antimicrobials within the first hour of recognized septic shock and hypotension is associated with an increase in survival rates. The critical nature of timing in the management of sepsis highlights the potential life-or-death consequences of delayed diagnosis.

- **Escalating Healthcare Costs:** The economic impact of sepsis is substantial and grows with delayed diagnosis and treatment [34]. Patients with sepsis require more resources, including extended stays in ICU, more complex and prolonged use of medications, and increased utilization of invasive procedures and technologies. These factors lead to higher direct healthcare costs. Furthermore, there are significant indirect costs to consider, such as lost productivity and long-term rehabilitation needs. Hospitals also face financial penalties related to high readmission rates, which are more common in sepsis survivors. Overall, the financial strain on healthcare systems can be severe, emphasizing the need for more effective early detection and management strategies to reduce the prevalence and severity of sepsis.

4 Procalcitonin as a Biomarker in Sepsis Prediction

4.1 Procalcitonin: A Foremost Indicator

PCT is a peptide precursor of the hormone calcitonin, the levels of which rise significantly in response to bacterial infection. Unlike other biomarkers such as CRP or Interlekin 6 (IL6) , which are elevated in a wide range of inflammatory states, PCT offers superior specificity to bacterial infections [39]. The clinical utility of PCT has been recognized in its ability to differentiate bacterial infections from other causes of inflammation. In clinical settings, PCT levels begin to rise 3-6 hours after a bacterial infection and peak at 12-48 hours, which is crucial for the timely diagnosis of sepsis. Elevated PCT levels are associated with a broad spectrum of bacterial infections, ranging from mild and localized to severe and systemic infections. This responsiveness makes PCT an indispensable tool in the early detection and management of sepsis, providing a crucial time window for intervention before severe sepsis or septic shock can occur.

The sensitivity of PCT as an indicator of sepsis was exemplified in a study conducted at a tertiary care center in Bangalore, which reported a PCT sensitivity of 94% among patients presenting with varying stages of sepsis [66]. This study further demonstrated that PCT levels are significantly associated with the Sepsis-related Organ Failure Assessment (SOFA) scores, indicating its relevance in real-time clinical assessment and its potential to guide therapeutic decisions effectively.

4.2 Limitations of Procalcitonin

While PCT offers specific benefits in diagnosing and managing sepsis, its limitations become more evident when compared with more commonly used tests such as the CBC, particularly in terms of cost and turnaround time.

- **Cost Comparison:** The cost of a PCT test in the United States can range significantly, from approximately $35 [31] to as high as $383 [21] depending on the setting and specifics of the test scenario, which is considerably higher than a CBC. A CBC test, widely utilized to assess overall health and detect a variety of disorders including infection and anemia, is much cheaper, typically costing between $10 to $30 per test depending on the healthcare setting. This stark difference in cost makes CBC a more feasible option for routine screening and initial diagnostic assessment in various clinical scenarios, particularly in resource-constrained environments.

- **Turnaround Time:** In terms of turnaround time, CBC tests also have an advantage. Results from a CBC can often be obtained in as little as a few minutes to an hour, especially with automated counters available in most modern laboratories. This rapid turnaround is crucial in emergency settings where quick decision-making is essential. In contrast, PCT results take anywhere from 2 to 4 hours. This delay could be critical in situations where immediate intervention is necessary.

5 Machine Learning in Sepsis Prediction

5.1 Machine Learning and Its Role in Healthcare

Machine Learning (ML) is a subset of artificial intelligence (AI) that focuses on developing algorithms capable of learning patterns and making decisions based on data. Unlike traditional programming, where explicit rules and instructions are provided, machine learning models improve their performance over time through exposure to more data, enhancing their predictive accuracy and adaptability. This process of learning from data allows ML to perform tasks that were once thought to be exclusively the domain of human expertise, such as speech recognition, image processing, and now, complex medical diagnostics.

In the hierarchy of AI technologies, ML lies in the intersection between data science and AI. AI encompasses a broader scope, including rule-based systems, natural language processing, and robotics, while ML is specifically centered on data-driven decision-making.

5.2 Machine Learning Approaches Used for Sepsis Detection

ML models have demonstrated substantial efficacy in the early detection of sepsis, often outperforming traditional diagnostic tools. A variety of algorithms and evaluation metrics have been harnessed to process and analyze clinical data, offering promising results in predicting the onset of sepsis. For instance, studies utilizing Electronic Health Records

(EHR) [1] data report Area Under the Receiver Operating Characteristic (AUROC) scores exceeding 0.90, indicating robust predictive accuracy [35]. These models excel in integrating and scrutinizing large and intricate datasets, identifying nuanced patterns indicative of the early stages of sepsis, which may elude traditional diagnostic methods.

5.2.1 Overview of Machine Learning Algorithms

The Table 2.1 will explore a comprehensive list of algorithms that have shown promise in the context of sepsis prediction, each characterized by its approach to learning from data and its performance metrics.

Table 2.1: Machine Learning Models Used in Detection of Sepsis

Abbreviation	**Full Name**	**Definition**
KNN [58]	K-Nearest Neighbors	A non-parametric method used for classification and regression; an object is classified by a majority vote of its neighbors, according to the k-nearest instances.
LR [59]	Logistic Regression	A statistical model that estimates the probability of a binary outcome, commonly used in medical diagnostics for its simplicity and effectiveness.
NB [49]	Naïve Bayes	A simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.
SVM [56]	Support Vector Machine	A powerful classification technique that finds a hyperplane in an N-dimensional space to distinctly classify data points, effective in high-dimensional spaces.
XGBoost [69]	eXtreme Gradient Boosting	An implementation of gradient boosted decision trees designed for speed and performance.

[1] Electronic Health Records (EHR) are digital versions of patients' medical histories, maintained over time by the health care provider. They contain data including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight

Abbreviation	Full Name	Definition
ANN [55]	Artificial Neural Network	A computational model based on the structure and functions of biological neural networks, which consists of an interconnected group of artificial neurons.
LSTM [30]	Long Short-Term Memory	A type of recurrent neural network capable of learning order dependence in sequence prediction problems, commonly used in deep learning.
RF [51]	Random Forest	An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of the individual trees.
AdaBoost [47]	Adaptive Boosting	An ensemble boosting classifier that combines multiple weak classifiers to increase the accuracy of classifiers, adjusting weights of incorrectly classified instances.
RUSBoost [27]	Random Under Sampling Boosting	A variation of the AdaBoost algorithm that incorporates random under sampling to address class imbalance in the training dataset, particularly effective for large imbalances.
Stacking [54]	Stacking	An ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor to improve prediction accuracy.

5.2.2 Evaluation Metrics for Predictive Models

The accuracy and reliability of machine learning models in predicting sepsis are primarily assessed using various statistical metrics. The Table 4.3 introduces the key evaluation metrics used in studies to measure the predictive power of the machine learning algorithms.

Table 2.2: Evaluation Metrics of Machine Learning Models

Metric Name	**Definition**	**Mathematical Equation**
Accuracy [57]	The proportion of true results (both true positives and true negatives) among the total number of cases examined.	$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total Cases}}$
Precision [50]	The ratio of true positives to the sum of true and false positives, indicating the accuracy of positive predictions.	$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$
Recall (Sensitivity) [52]	The ratio of true positives to the sum of true positives and false negatives, showing how well the model can identify positive cases.	$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$
F1 Score [48]	The harmonic mean of precision and recall, providing a balance between the two metrics for situations where an equal trade-off is necessary.	$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$
AUROC [53]	A performance metric for binary classification problems, AUROC represents the degree of separability between classes achieved by the model. It is the area under the ROC curve, a graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. The AUROC quantifies the overall ability of the model to correctly classify positive and negative instances across all possible thresholds, effectively summarizing the trade-off between true positive rate (sensitivity) and false positive rate. A higher AUROC value indicates better model performance, with a score of 1.0 representing perfect classification and 0.5 indicating no discriminative ability.	$\text{AUROC} = \int_0^1 \text{TPR}(t)\, dt$ where $\text{TPR}(t)$ is the true positive rate at threshold t

6 Literature Review

In this section, we explore the body of existing research related to the application of ML in the early detection, diagnosis, and management of sepsis. Recent years have witnessed AI emerging as a pivotal technology in healthcare. With the exponential growth in healthcare data and computational power, AI and ML models have shown promise in enhancing sepsis outcomes.

The Table 2.3 and the Table 2.4 are dedicated to a comprehensive comparison of existing research studies on the use of AI for the early detection, diagnosis, and management of sepsis. By comparing the results and approaches across different research works, we aim to discern patterns of success and limitations within the current landscape.

Table 2.3: Summary of Research Studies - Part 1

Paper	**Research Objective**	**Dataset & Features**	**ML model(s)**	**Key Findings and Results**	**Limitations**
[35]	Detection of sepsis in ICU patients.	MIMIC-III and UCSF datasets with vital signs as features.	XGBoost	AUROC : 0.88	Limited to the use of vital signs.
[7]	Compare the effectiveness of different ML models in predicting sepsis.	Data from Rabin Medical Center including only vital signs.	ANN, SVM and LR	All models reached an AUROC of 0.88	Very limited set of features.
[29]	Predict sepsis using time-series data from ICU patients.	MIMIC-II data including demographics, labs, and vitals.	LSTM	AUROC : 0.99	The model may not generalize well across different ICU settings.
[67]	Develop an ML model to predict sepsis early in ICU patients.	EHR data from 4,449 infected ICU patients with 55 features including clinical and laboratory parameters.	RF	AUROC: 0.91, Sensitivity : 87% and specificity : 89%,	Single-center data source, which may not generalize across different populations.
[9]	Develop and compare ML models for the early prediction of sepsis in ICU patients.	MIMIC-III dataset	RF, SVM, KNN, ANN, NB, AdaBoost, Stacking, and XGBoost and Ensemble techniques	Models that involve stacking techniques and XGBoost, significantly outperformed traditional models like SOFA and qSOFA scores.	The study's reliance on a single dataset may limit generalizability. The computational complexity of ensemble models also requires substantial processing power.

Table 2.4: Summary of Research Studies - Part 2

Paper	Research Objective	Dataset & Features	Machine Learning model(s)	Key Findings and Results	Limitations
[64]	Develop a ML model to predict sepsis using CBC diagnostics.	Non-ICU data from a German tertiary care center. Two external validations included data from another German center and MIMIC-IV. Features included patient age, sex, and CBC parameters: Hemoglobin, Platelets, Mean Corpuscular Volume, White and Red blood cells.	RUSBoost.	AUROC : 0.872, External validations AUROCs : 0.805 and 0.845 for the two additional datasets. The model including PCT showed an AUROC of 0.857.	The model excludes surgical ICU admissions and SIRS cases.

The literature review reveals that the efficacy of machine learning models in predicting sepsis is significantly influenced by the choice of input features, with vital signs, laboratory values, and demographic data being the most critical [10]. Ensemble methods like Random Forest and XGBoost, along with deep learning approaches such as LSTM, stand out due to their ability to handle complex datasets and identify early sepsis indicators. Evaluation metrics such as AUROC is crucial for assessing model performance.

Conclusion

This chapter has outlined the complex and multifaceted nature of sepsis, from its initial onset to its most severe stages. By exploring the clinical significance, progression, and challenges associated with sepsis, along with a review of the latest scientific advancements, we have underscored the urgent need for more accurate and timely diagnostic approaches.

3

Key Features Selection

Introduction

This chapter delineates the comprehensive methodology adopted to explore and analyze datasets crucial for predicting sepsis. A detailed examination of the datasets sets the stage for a robust analysis, utilizing a variety of preprocessing and machine learning techniques to enhance the predictive accuracy for sepsis. This chapter also articulates the steps involved in feature selection, and the utilization of advanced analytical models, ensuring a systematic approach to understanding and leveraging the data for optimal results.

1 Dataset Overview

The dataset utilized in this research was collected from the Klinikum of Passau with a primary focus on facilitating the early prediction of sepsis. It encompasses 109,897 blood test results, capturing a comprehensive range of 26 different blood characteristics, alongside demographic information including age and sex. The blood tests were performed on hospitalized patients who, at the time of testing, exhibited no clinical symptoms of sepsis. These routine tests, part of standard clinical care, provide an invaluable dataset for early sepsis prediction, as they offer a window into the physiological state of patients before the onset of overt infection signs.

The patient population spans a broad age range, from 2 to 100 years old, including both male and female individuals. This demographic diversity is crucial, as it allows the model to account for the distinct physiological characteristics of pediatric, adult, and geriatric populations, as well as sex-specific variations. By utilizing routine data from asymptomatic patients, the dataset is instrumental in training predictive models aimed at detecting sepsis in its earliest stages.

2 Overview of Blood Features

The features include a spectrum of markers such as Procalcitonin (PCT) and Interleukin 6 (IL6), which are significant for indicating the severity of bacterial infections and inflammatory responses, respectively. Hematological parameters like Hematocrit (HK), Hemoglobin (HB), and various white blood cell counts provide essential clues about the body's response to infection and its capability to transport oxygen and manage immune responses. Advanced metrics like Mean Platelet Volume (MPV) and Red Cell Distribution Width (RDW-SD) offer further details on blood cell characteristics that may indicate pathological changes before clinical symptoms become apparent. The Table 3.1 lists all the blood features contained in the dataset.

Table 3.1: Dataset Features

Abbreviation	**Full Name**	**Definition**
BASOAB	Absolute Basophil Count	Counts basophils, the least common type of white blood cells.
BASOPH	Basophils Percentage	Shows the proportion of basophils in white blood cells.
EOSABS	Absolute Eosinophil Count	Measures eosinophils, indicating immune response to parasitic infections.
EOSINO	Eosinophils percentage	Measures the proportion of eosinophils, important in allergic and parasitic responses.
ERY	Erythrocytes count	Also known as red blood cells, crucial for oxygen transport.
GRANAB	Granulocytes Absolute Count	Shows the count of neutrophils in white blood cells.
GRANUL	Absolute Granulocytes count	Shows the count of neutrophils in white blood cells, crucial for fighting infections.
HB	Hemoglobin	A protein in red blood cells responsible for oxygen transport.
HK	Hematokrit	Measures the proportion of red blood cells in blood.
IG-C	Immature Granulocytes Count	Counts immature granulocytes, a marker of immune response to infection.
IG-P	Immature Granulocytes Percentage	Measures immature granulocytes in white blood cells, indicating bone marrow response.
IL6	Interleukin 6	A cytokine involved in inflammation and immune response.
LEUKO	Leukocytes count	White blood cells key in defending against infections; their count varies with immune response.
LYMABS	Absolute Lymphocyte Count	Measures the number of lymphocytes, a type of white blood cell significant in the immune system and fighting infections.

Table 3.1 Dataset Features

Abbreviation	**Full Name**	**Definition**
LYMPHO	Lymphocyte Percentage	Shows the proportion of lymphocytes in white blood cells.
MCH	Mean Corpuscular Hemoglobin	Measures average hemoglobin per red blood cell.
MCHC	Mean Corpuscular Hemoglobin Concentration	Indicates hemoglobin concentration in red blood cells.
MakroE	Macrocytes count	Larger than normal red blood cells.
MicroE	Microcytes count	Red blood cells smaller than normal.
MONABS	Absolute Monocyte Count	Indicates the number of monocytes, important for the immune system's chronic responses.
MONOZY	Absolute Monocytes count	Measures the number of monocytes, key in phagocytosis and immune response.
MPV	Mean Platelet Volume	Indicates the average size of platelets, with variations suggesting changes in platelet production.
MCV	Mean Corpuscular Volume	Measures the average size of red blood cells.
PCT	Procalcitonin	A biomarker indicating the severity of bacterial infection.
RDW-SD	Red Cell Distribution Width - Standard Deviation	Measures variation in red blood cell size.
THROMB	Thrombocytes / Platelets count	Cell fragments crucial for blood clotting.

3 Utilizing PCT Levels for Categorization

According to the findings by Sudhir et al. (2011) [66], PCT levels can be categorized into three significant groups, each influencing the diagnostic strategy for sepsis:

- **Category N (<0.1 ng/mL)**: Patients do not have sepsis.

- **Category P (>2 ng/mL)**: Patients are likely to have sepsis.
- **Category I (0.1 – 2 ng/mL)**: Patients are in an indeterminate state, and their sepsis status is uncertain.

In our dataset, PCT values are utilized to categorize patients into three groups, encoded as P (Sepsis-Positive), N (Sepsis-Negative), and I (Sepsis-Indeterminate). Since we do not have confirmed diagnoses of sepsis, these categorizations rely on PCT levels to label patients, aligning with clinical insights that recognize PCT as a reliable indicator of sepsis. This method not only streamlines the modeling process but also serves as an indirect validation of sepsis presence based on PCT levels.

To illustrate the PCT thresholds used for categorization, Figure 3.1 visualizes the distribution of PCT levels across the defined categories

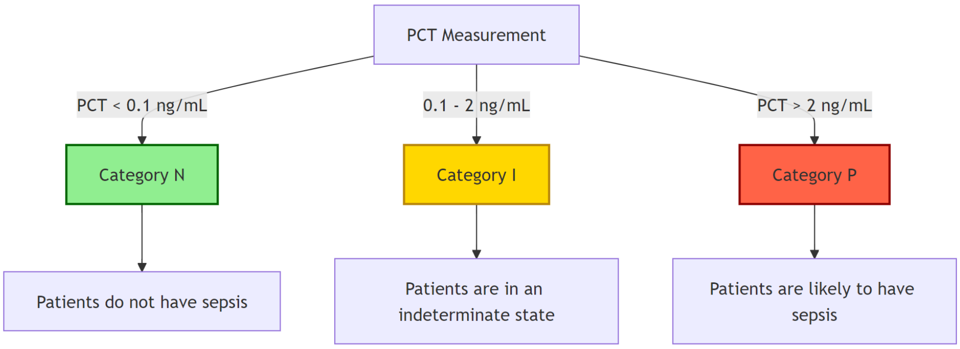

PCT Measurement
PCT < 0.1 ng/mL
0.1 - 2 ng/mL
PCT > 2 ng/mL
Category N
Category I
Category P
Patients do not have sepsis
Patients are in an indeterminate state
Patients are likely to have sepsis

Figure 3.1: PCT Thresholds

The distribution of the categories, as revealed by the analysis, is as follows:

- Sepsis-Negative (Category: N): 58779 cases (53.5%)
- Sepsis-Indeterminate (Category: I): 39974 cases (36.4%)
- Sepsis-Positive (Category: P): 11144 cases (10.1%)

The pie chart visualization of the data on the Figure 3.2 underscores the prevalence of cases classified under 'Sepsis-Negative', followed by 'Sepsis-Indeterminate', with 'Sepsis-Positive' constituting a relatively small fraction.

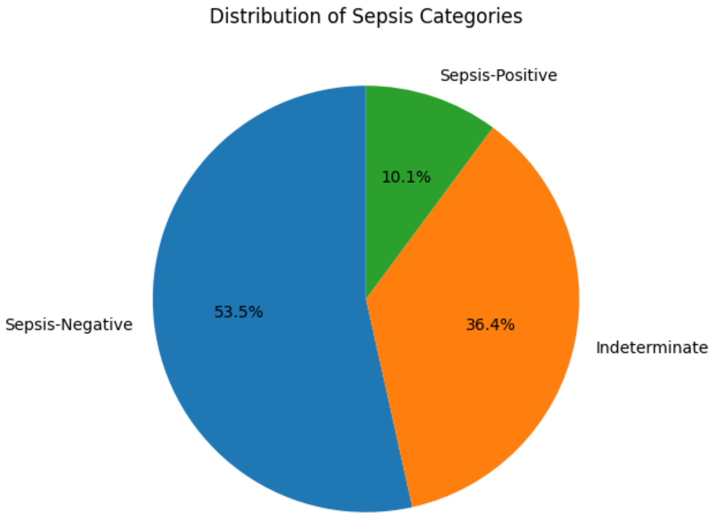


Figure 3.2: Proportion of Samples

This imbalanced distribution is characteristic of real-world clinical datasets, where the incidence of confirmed cases of medical conditions like sepsis can be relatively low. The significant proportion of the 'Indeterminate' category emphasizes the diagnostic ambiguity often encountered in clinical settings.

4 Data Visualization

4.1 Age-Related Variations in Procalcitonin Levels

The scatter plot provided in Figure 3.3 clearly illustrates the distribution of PCT levels across various age groups. Horizontal reference lines set at PCT levels of 0.1 and 2 allow us to evaluate the spread of elevated PCT levels. Upon analyzing the data, we observe substantial variations in PCT levels across different age demographics. Crucially, levels exceeding the threshold of 2 are frequently seen among older adults.

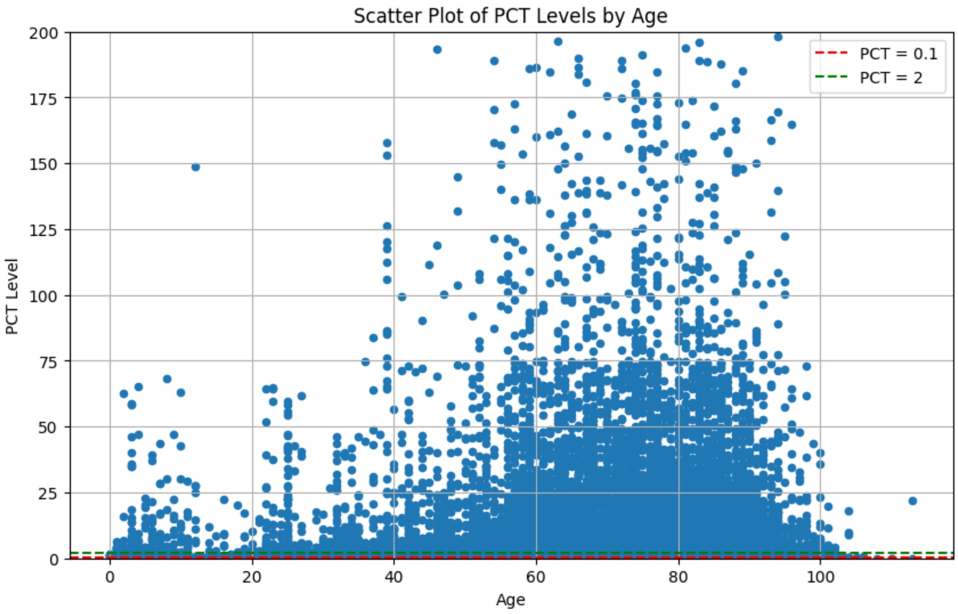


Figure 3.3: Scatter Plot of PCT Levels by Age

The box plot of age analysis for individuals labeled as Sepsis-Positive in Figure 3.4 reveals that within this subgroup: 25% are younger than 63 years, the median age in this category is 74 years and the third quartile is observed at 82 years. The mean age of approximately 70.7 years. This analysis clearly indicates that the majority of individuals labeled as Sepsis-Positive are elderly.

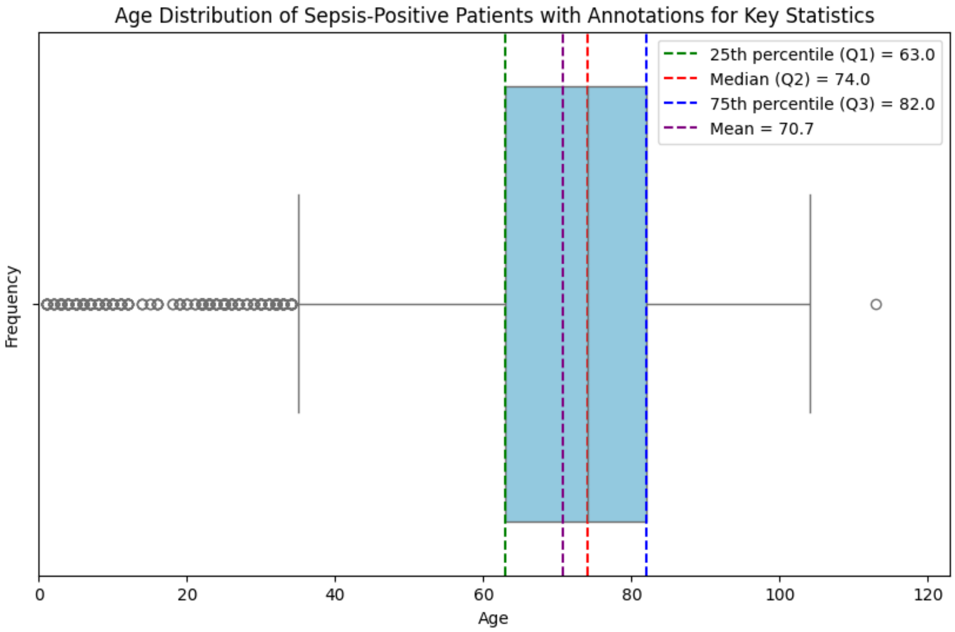


Figure 3.4: Age Distribution of Positive Patients

4.2 Sex Distribution in Patient Data

The demographic breakdown of the dataset was closely examined to ensure a balanced representation of gender, which is crucial for minimizing bias in the model's predictions. The dataset is composed of patient entries classified by sex, with a nearly equal distribution between males and females. With males constituting 53.4% and females making up the remaining 46.6%. This nearly balanced distribution is advantageous as it allows for the development of a model that is not skewed towards any particular gender, thereby enhancing the generalizability and fairness of the predictive analysis.

5 Key Features Selection

Feature selection is a fundamental step in our study as it directly influences the model's ability to generalize well to new data while avoiding overfitting. Overfitting occurs when a model is too complex, capturing noise instead of representing the underlying dataset distribution [25]. To mitigate this risk, we employ various analytical techniques to determine the most significant features that contribute to accurately predicting sepsis.

The objective of our feature selection process is to determine which features are most impactful in diagnosing sepsis while eliminating those that contribute little to predictive performance. To achieve this, we utilize a range of robust analytical models, each contributing insights into feature relevance and efficacy.

All the analysis was confined to data explicitly labeled as 'P' (Sepsis Positive) and 'N' (Sepsis Negative), with the objective of determining distinct patterns between septic and non-septic patients. This selective focus ensures that indeterminate states, which do not add value to the decision made by the algorithm, are excluded from the analysis.

5.1 Clustering Analysis

In our study, we employed the k-means clustering technique as an initial step in our exploratory data analysis. This method categorizes the dataset into two distinct clusters based on the similarity of patient blood count profiles, helping to identify patterns that guide the development of more effective diagnostic models.

To facilitate this analysis, we first standardized the dataset using the StandardScaler. Standardization involves rescaling each feature so that it has a mean (average) of zero and a standard deviation of one, mathematically represented as $z = \frac{(x-\mu)}{\sigma}$, where x is the original value, μ is the mean of the feature, and σ is the standard deviation. This process is crucial as it ensures that the clustering algorithm evaluates the features on a uniform scale, thereby mitigating the influence of outlier values and differing scales across the features.

Following the clustering, we utilized Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. The primary purpose of PCA in our study is to transform the high-dimensional data into a two-dimensional space, making it more amenable to visual interpretation. By reducing the data to two principal components, we provide a visual simplification that retains the most significant variance in the data, allowing for an intuitive graphical representation of the clusters.

Figure 3.5 presents the results of the PCA conducted after clustering. The left panel, titled *PCA Cluster Visualization by k-Means Clustering*, displays the dataset partitioned into two clusters, each represented by a different shade. The right panel, titled *PCA Cluster Visualization by Original Labels*, color-codes the data points according to their sepsis status: 'P' for sepsis Positive shown in red, and 'N' for sepsis Negative depicted in blue.

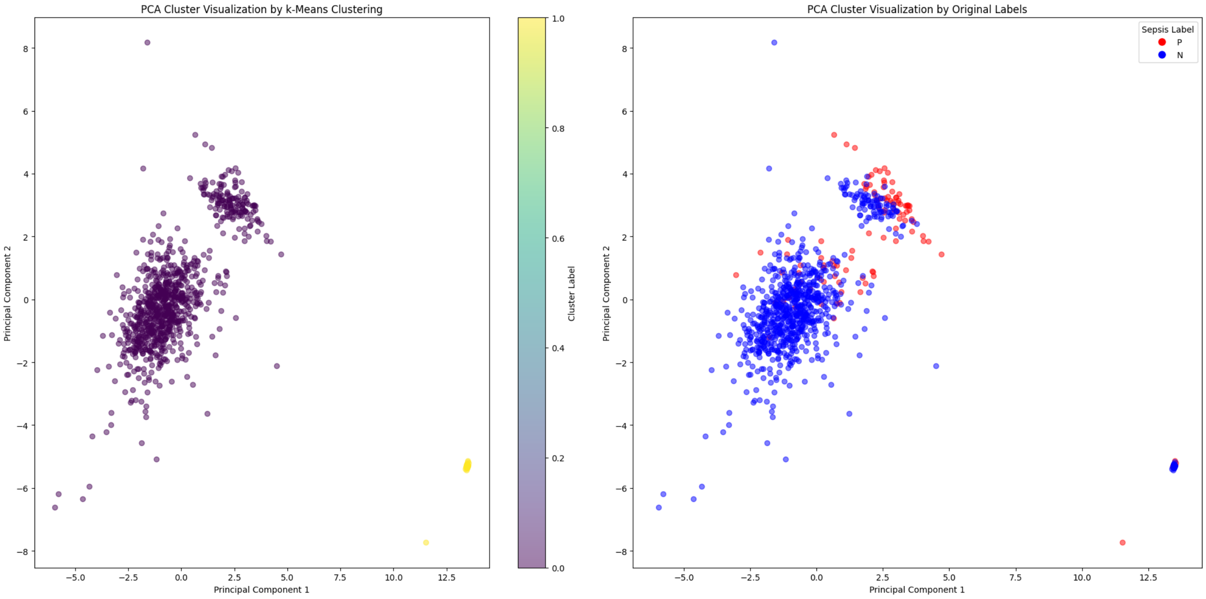


Figure 3.5: Cluster Analysis

The Clustering Visualization by k-Means Clustering displays a gradient of clusters without clear separation, complicating the identification of distinct groups representing septic and non-septic cases. Furthermore, the cluster visualization color-codes the data points as red for sepsis positive ('P') and blue for sepsis negative ('N'), where ideally we would expect to see a clear separation between the red and blue points if the clustering effectively captured the underlying patterns specific to sepsis outcomes. This variance is indicative of the heterogeneous nature of sepsis, where multiple underlying physiological processes and patient-specific factors contribute to the disease presentation.

These initial findings underscore the potential limitations of relying solely on linear dimensionality reduction techniques and linear clustering methods like PCA and k-means for complex medical diagnostics such as sepsis. The results suggest a need for more sophisticated modeling approaches that can handle non-linear relationships and interactions within the data.

5.2 Correlation Analysis Between PCT and Other Blood Features

Following the inconclusive results from the clustering analysis, we adapted our approach to further understand the correlation of blood feature and PCT.

5.2.1 Linear Correlation: Pearson's Correlation

Linear correlation measures the strength and direction of a linear relationship between two variables. In the study of sepsis, assessing how changes in PCT levels linearly correlate with variations in other blood biomarkers provides preliminary insights into potential diagnostic markers. Due to the inherent complexity and noise in biological data, identifying subtle yet potentially significant patterns is crucial. Thus, the threshold for correlation magnitude was intentionally set at a modest 0.2, both positive and negative. This value is strategically chosen to identify potential, though not necessarily strong, correlations that could indicate biologically significant relationships.

This modest threshold is particularly useful in the context of medical data analysis, where biological relevance often supersedes strong statistical signals. It allows for the inclusion of biomarkers that show even weak correlations with PCT levels, which may still play a significant role in the inflammatory processes characteristic of sepsis. By setting this threshold, we ensure that no potentially important markers are overlooked, thereby aiding in the preliminary identification of relationships that might warrant further investigation.

The results of our correlation analysis are visually represented in Figure 3.6, where vertical lines at 0.2 and -0.2 clearly delineate the biomarkers that show correlation magnitudes beyond this set threshold.

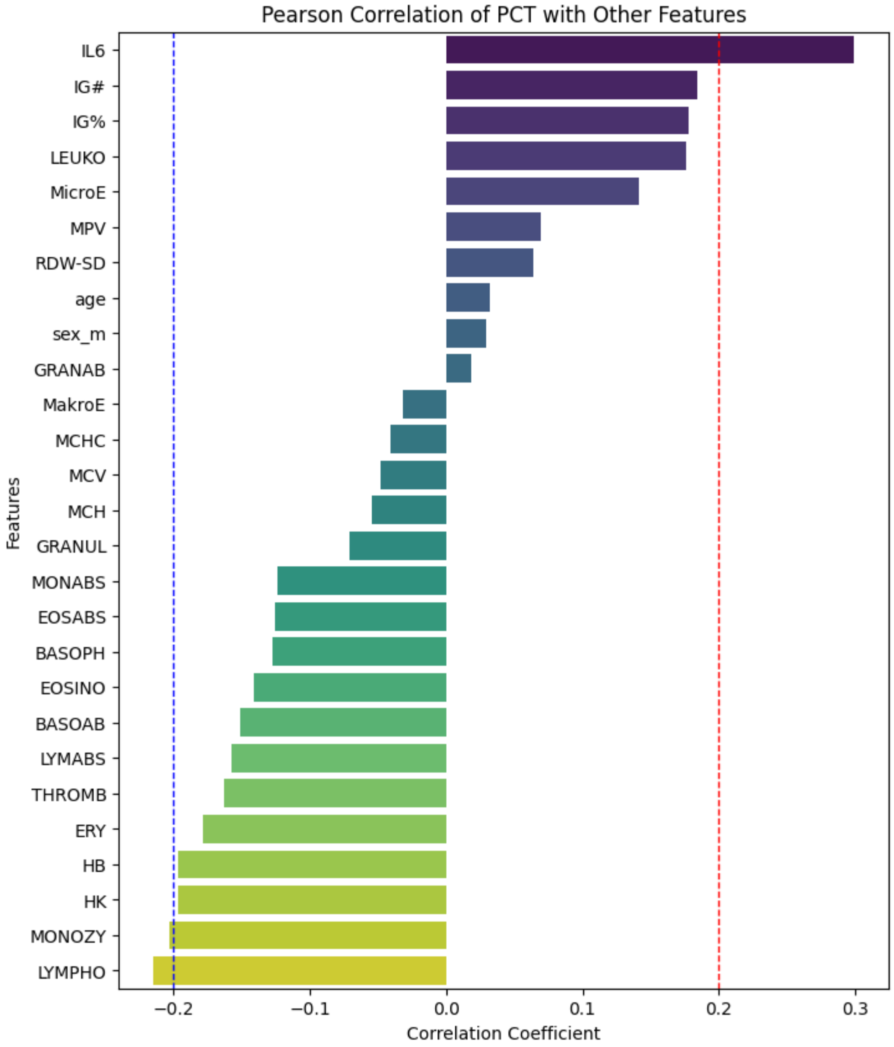


Figure 3.6: Pearson's Correlation of PCT with Other Blood Features

Notably, only a few features exhibit correlation coefficients beyond these bounds:

- **Positive Correlation:** Interleukin 6 with a coefficient of 0.298.
- **Negative Correlation:** Lymphocytes and Monocytes Absolute Count with coefficients of -0.214 and -0.202, respectively.

Although these identified correlations are valuable, it is important to recognize the limitations of linear correlation analysis. This method does not capture non-linear relationships, which may be significant in the complex biological interactions inherent to sepsis. Thus, relying solely on Pearson's correlation might result in overlooking other crucial biomarkers that interact with PCT in a non-linear manner.

5.2.2 Spearman's correlation

Spearman's correlation is adept at detecting monotonic relationships that are not necessarily linear but show consistent trends in either increasing or decreasing together. This approach is particularly beneficial in uncovering the complex and nonlinear interactions that are prevalent in the biological pathways. The results of our Spearman's correlation analysis are illustrated in Figure 3.7, where the vertical lines at 0.2 and -0.2 help delineate the biomarkers with significant monotonic relationships with PCT levels.

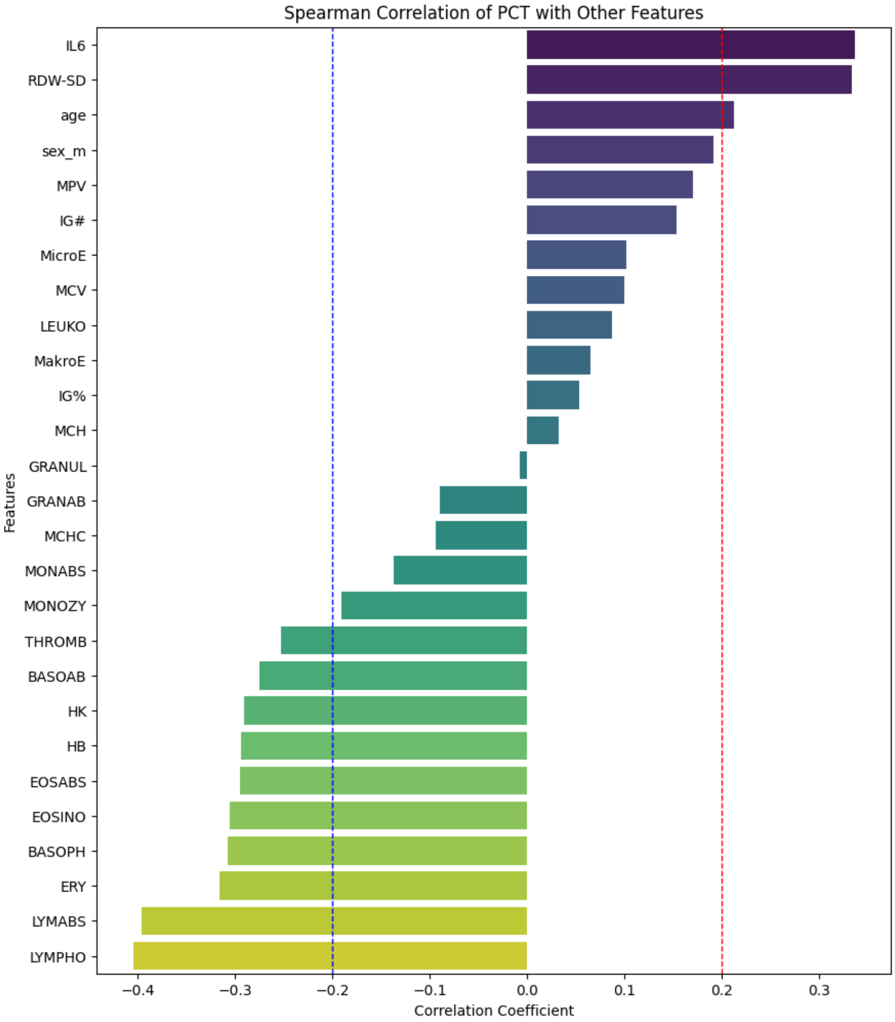


Figure 3.7: Spearman Correlation of PCT with Other Blood Features

- **Significant Monotonic Positive Relationships:** Biomarkers such as Interleukin 6, Red Cell Distribution Width-Standard Deviation and Age which exhibited a Spearman's correlation coefficient significantly above 0.2, suggest a robust, consistent increase alongside PCT levels.
- **Significant Monotonic Negative Relationships:** Similarly, markers like Lymphocytes pecentage, Lymphocytes Absolute Count, Erythrocytes, Basophils percentage, Eosinophils percentage, Eosinophils Absolute, Hemoglobin, Hematocrit, Basophils Absolute and Thrombocytes, with coefficients below -0.2, indicate a consistent decrease relative to rising PCT levels.

These findings, illustrate how Spearman's correlation has enabled us to effectively segregate features that are correlated with PCT from those that are not especially when dealing with variables that interact in non-linear ways.

5.3 Box Plots Method

To explore the variations and discern any significant differences in the dispersion of values among individuals across different sepsis categories ('P', 'N', and 'I'), we employed box plots for a detailed visual examination of each feature across these categories.

Box plots were chosen as the primary analytical tool due to their effectiveness in depicting the distribution of data. These plots facilitate an understanding of the median, quartiles, and the presence of outliers within each category. This method is particularly beneficial for identifying whether individuals in the 'P' category show higher or lower ranges of certain features compared to those in the 'N' category, and how the 'I' cases compare to both.

This approach aligns with the exploratory nature of the study, where the primary goal is to sift through complex, high-dimensional data to pinpoint markers that warrant further investigation. As outlined in statistical methodologies, the use of box plots to compare distributions provides a clear, visual representation of the central tendency and variability within data, making it easier to spot these distinctions.

The analysis was structured into two key phases:

1. **Data Segregation:** The dataset was sorted into three distinct groups based on the sepsis labels—'P', 'N', and 'I'.

2. **Visualization:** A box plot was generated for each feature. In these plots, the x-axis represented the sepsis category, and the y-axis depicted the range of values for the feature. This allowed for a comparative analysis across categories.

This detailed examination yielded several critical insights:

- **Identifying Key Features:** Features where the ‘P’ category demonstrated significantly higher or lower values compared to the ‘N’ category were flagged as potentially crucial for predicting sepsis.
- **Utility Assessment:** Conversely, features exhibiting considerable overlap across categories were deemed to have lower discriminative power, thus considered less useful for predictive modeling.

To enhance the interpretability of our sepsis prediction model, we introduce a decision threshold criterion based on the non-overlapping interquartile ranges (IQR) of the box plots. This criterion helps in distinguishing between features that provide clear differentiation across sepsis categories from those that do not. Specifically, a feature is considered highly discriminative if at least 50% of its data range (IQR) for one category does not overlap with 50% of the data range of another category, as shown in Figure 3.8. Conversely, features where the IQRs overlap significantly across categories, as in Figure 3.9, are considered less useful for predictive modeling.

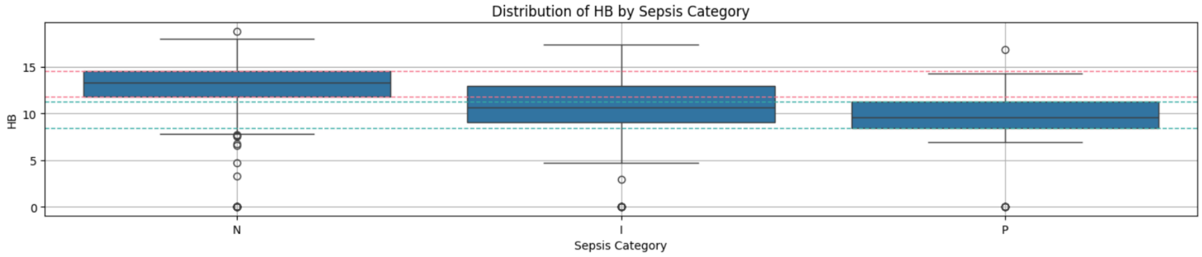


Figure 3.8: Example of distribution of HB

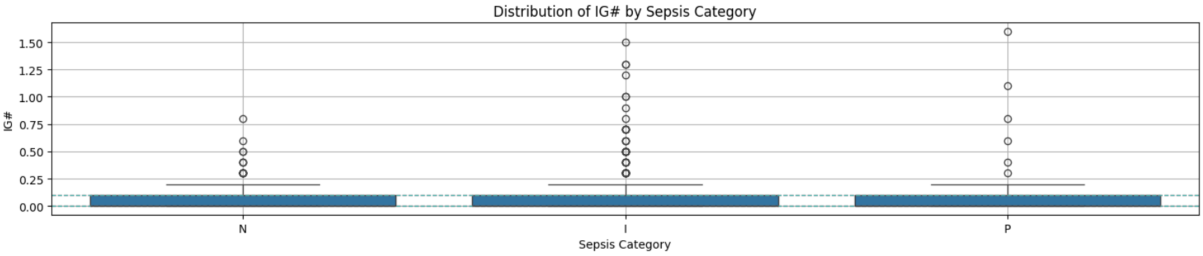


Figure 3.9: Example of distribution of IG-C

Based on our established criterion, we categorized the features based on their interquartile range (IQR) non-overlap, distinguishing them into groups indicative of their discriminative power. The features demonstrating minimal IQR overlap, and thus providing clear differentiation across sepsis categories, were identified as important for predictive modeling. These include Hematocrit, Lymphocytes Absolute Count, Monocytes Absolute Count, Hemoglobin, Lymphocytes percentage, Erythrocytes, Eosinophils Absolute Count, Eosinophils percentage, Leukocytes, Monocytes Absolute Count and Basophils Percentage.

By focusing exclusively on these clinically relevant biomarkers, our approach systematically emphasizes the most significant indicators for diagnosing sepsis, enhancing the model's effectiveness by omitting features with significant overlap that are deemed less useful for predictive analysis.

5.4 Random Forest Analysis for Sepsis Prediction

In our pursuit to enhance the predictive analytics for sepsis, we employed a Random Forest classifier, recognized for its robustness in handling tabular data and its ability to manage complex feature interactions. Random Forests excel in this environment due to their inherent feature selection capabilities through importance scoring, making them ideally suited for our dataset [37].

The dataset comprised multiple variables from which we excluded the PCT levels to delineate our features (`X`) and target (`y`) variables. The target variable was binary encoded where 'P' (Sepsis-Positive) and 'N' (Sepsis-negative) were mapped to 1 and 0, respectively. The dataset was partitioned into training and testing sets with an 80-20 split, ensuring a random and reproducible allocation of data using a seed (`random_state=42`). A `RandomForestClassifier` with 100 estimators was trained on the training set. We extracted and analyzed the feature importance scores generated by the classifier, providing insights into the most influential variables contributing to sepsis detection. The Figure 3.10 illustrates the flowchat of the RandomForestClassifier.

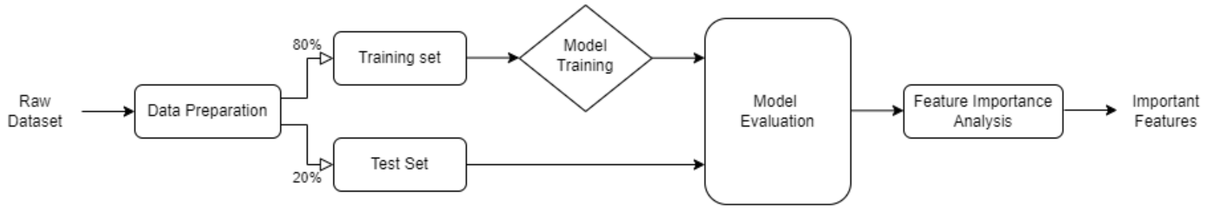


Figure 3.10: Flowchart of the Random Forest Model

The analysis was conducted on the feature importance scores derived from the RandomForestClassifier to identify the most critical predictors of sepsis in our dataset. The importance score, which quantifies the contribution of each feature to the predictive accuracy of the model, was calculated during the training phase. A threshold line at an importance score of 0.03 was introduced to delineate the features with the highest impact. Setting this threshold helps in distinguishing features that significantly contribute to the model's predictive power from those that may only add noise. A threshold of 0.03 ensures that only features with a non-negligible impact on the model's performance are considered.

To visually represent the distribution of feature importance and highlight the most influential factors, we plotted these scores using a horizontal bar chart on Figure 3.11 in descending order.

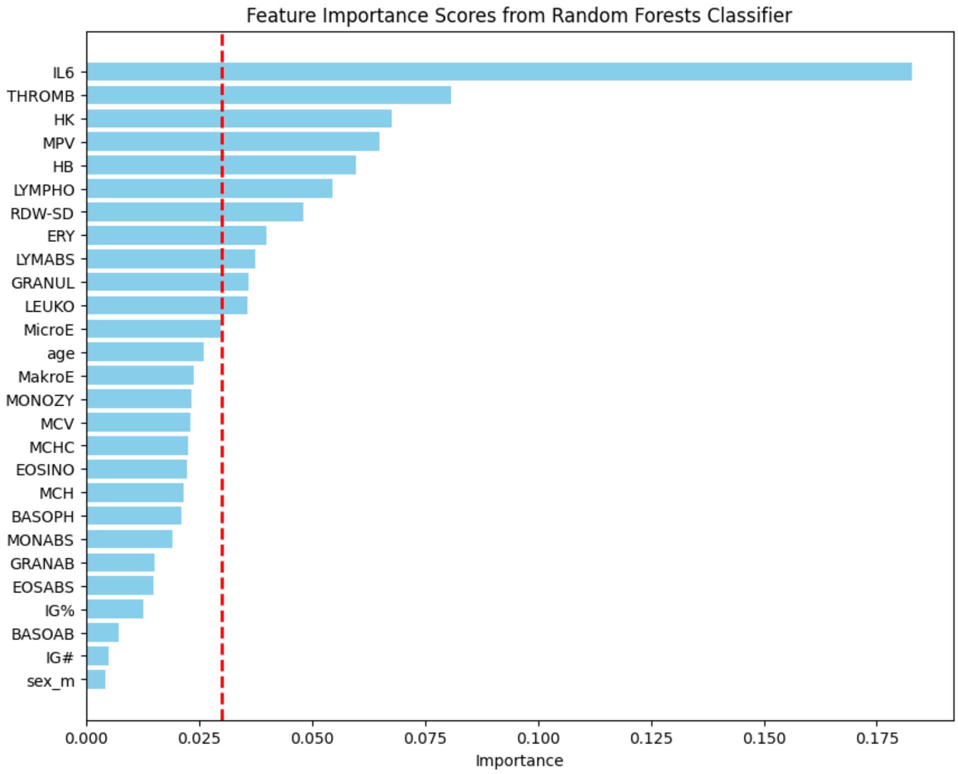


Figure 3.11: Feature Importance Scores with Random Forest Model

The identified featuress such as Interleukin 6, Thrombocytes Count, Hematocrit, Mean Platelet Volume, Hemoglobin, Lymphocytes Percentage, Red Cell Distribution Width-Standard Deviation, Erythrocytes Count, Lymphocytes Absolute Count, Granulocytes Percentage and Leukocytes Count emerged as the top predictors, indicating their significant roles in distinguishing between sepsis-positive and sepsis-negative cases.

5.5 Evaluation of Feature Importance Using eXtreme Gradient Boosting Classifier

As part of our ongoing effort to refine predictive models for sepsis,the eXtreme Gradient Boosting (XGBoost) classifier, a decision-tree-based ensemble machine learning algorithm was deployed to analyze the impact of various features on predicting the occurrence of sepsis. Known for its outstanding performance and efficiency in classification tasks [2], XGBoost is particularly adept at handling diverse and imbalanced datasets, making it an excellent tool for our sepsis prediction research.

The dataset utilized retained the same variables as previously described, excluding PCT levels. The target variable was encoded in a binary format where ‘P’ and ‘N’ corresponded to 1 and 0, respectively. This setup facilitates the direct application of binary logistic regression as the objective function in our XGBoost model.

The data was divided into training and testing sets with an 80-20 split, and the partitioning was made reproducible through the use of a consistent seed (`random_state=42`), ensuring that results could be reliably compared across models. An XGBoost classifier was configured with 100 estimators and the binary logistic regression objective to specifically address the binary nature of our target variable. The model was trained on the training dataset.

Post training, the model’s performance was evaluated on the testing set. To elucidate the importance of each feature in the model, we utilized XGBoost’s built-in functionality to calculate and display feature importance based on three metrics: weight, gain, and coverage [40]. Each metric offers a different perspective on the data:

- **Weight:** refers to the number of times a feature is used to split the data across all trees. A higher weight indicates that the feature is more frequently used for making crucial decisions.

- **Gain:** Gain measures the average "improvement" in accuracy brought by a feature to the branches it is used in.
- **Coverage:** Coverage measures the average number of observations affected by the splits where a particular feature is used. This metric considers how much data is involved when a feature is utilized in tree branches.

These metrics were visually represented in a series of plots, allowing for an intuitive understanding of feature contributions to the model. Figure 3.12 presents these plots, demonstrating the comparative importance of the top six features as per weight, gain, and coverage. The visualization was performed using F1 scores on the x-axis, selected for its balanced consideration of both precision and recall, making it particularly suitable for evaluating model performance in the context of imbalanced datasets.

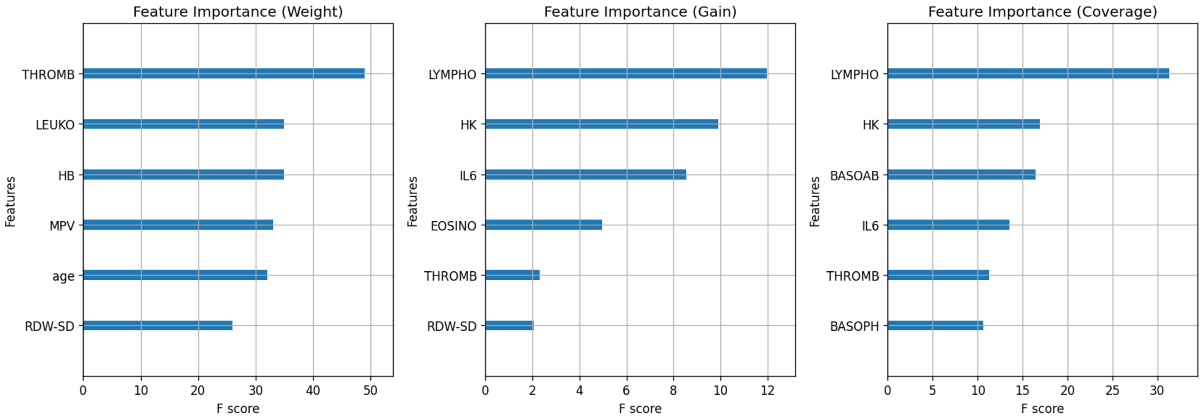


Figure 3.12: Feature Importance Scores with XGBoost Model

Upon this analysis of the XGBoost classifier, it becomes evident how certain features significantly influence the model's decision-making process in predicting sepsis. Features such as Thrombocytes Count, Leukocytes Count, Hemoglobin, Mean Platelet Volume, age and Red Cell Distribution Width-Standard Deviation are used extensively across decision trees, indicating their substantial role in the model's decisions through frequent usage in data splits. Conversely, features like Lymphocytes Percentage, Hematocrit, Interleukin 6, Eosinophils Percentage, Thrombocytes Count and Red Cell Distribution Width-Standard Deviation markedly enhance the model's performance. Their ability to significantly improve prediction accuracy demonstrates their effectiveness. Moreover, Lymphocytes Percentage, Hematocrit, Basophils Absolute Count, Interleukin 6, Thrombocytes Count and Basophils Percentage also play a critical role in generalizing the model to a broader dataset, which is vital for its application across diverse clinical settings.

5.6 Feedforward Neural Network and Permutation Importance

A Feedforward Neural Network (FNN) was developed using TensorFlow and Keras libraries to predict sepsis from clinical data. The data preparation phase involved standardizing the dataset to ensure uniformity in the measurement scales, which is crucial for neural network performance. After standardization, the dataset was divided into two distinct sets: 80% of the data for training and 20% of the data for testing.

The architecture of the neural network utilized in this study consists of multiple dense layers, which are fully connected layers known for their effectiveness in pattern detection within deep learning frameworks. To combat the common issue of overfitting—where a model learns the training data's noise and fails to generalize to unseen data—dropout regularization was employed. Dropout works by randomly deactivating a subset of neurons during the training phase, thereby forcing the model to learn more robust features that are not overly reliant on any individual neuron's activation.

In our specific model setup, the neural network was configured with dual dropout layers interspersed between dense layers. Each dropout layer was set to deactivate 50% of the neurons. This substantial dropout rate was chosen to ensure robust feature learning and to prevent overdependence on particular neurons, which is crucial given the complexity and potential overfitting risks associated with our dataset. The network's architecture included an initial dense layer of 128 neurons, followed by a dropout layer, then a second dense layer of 64 neurons, and another dropout layer, culminating in a single neuron output layer for binary classification. Such a configuration of dropout layers at a rate of 0.5 was critical in enhancing the model's ability to generalize well on new, unseen data, as indicated by the improved performance metrics on the test set.

Training was executed over 50 epochs, with the possibility of adjustment if the trends in training and validation losses indicated the need for longer or shorter training durations. This flexibility in epochs allows for optimized learning where the model has sufficient iterations to learn from the data without memorizing it.

Post-training, we employed permutation importance to elucidate the significance of each feature in predicting sepsis. This analytical method entails the following steps:

- Each feature within the test dataset is systematically shuffled, disrupting the order of data points solely for that specific feature.

- The performance of the model is then evaluated to quantify the impact of this perturbation. Specifically, we measure the degradation in the model's accuracy, which is a direct indicator of the feature's importance.
- A substantial decline in model accuracy upon shuffling a feature suggests its critical role in the accurate prediction of sepsis.

The outcomes are visualized in a horizontal bar chart of Figure 3.13, where each bar represents the feature's importance based on the observed decrease in model accuracy. a threshold line at 0.01 on the X-axis serves as a critical demarcation.

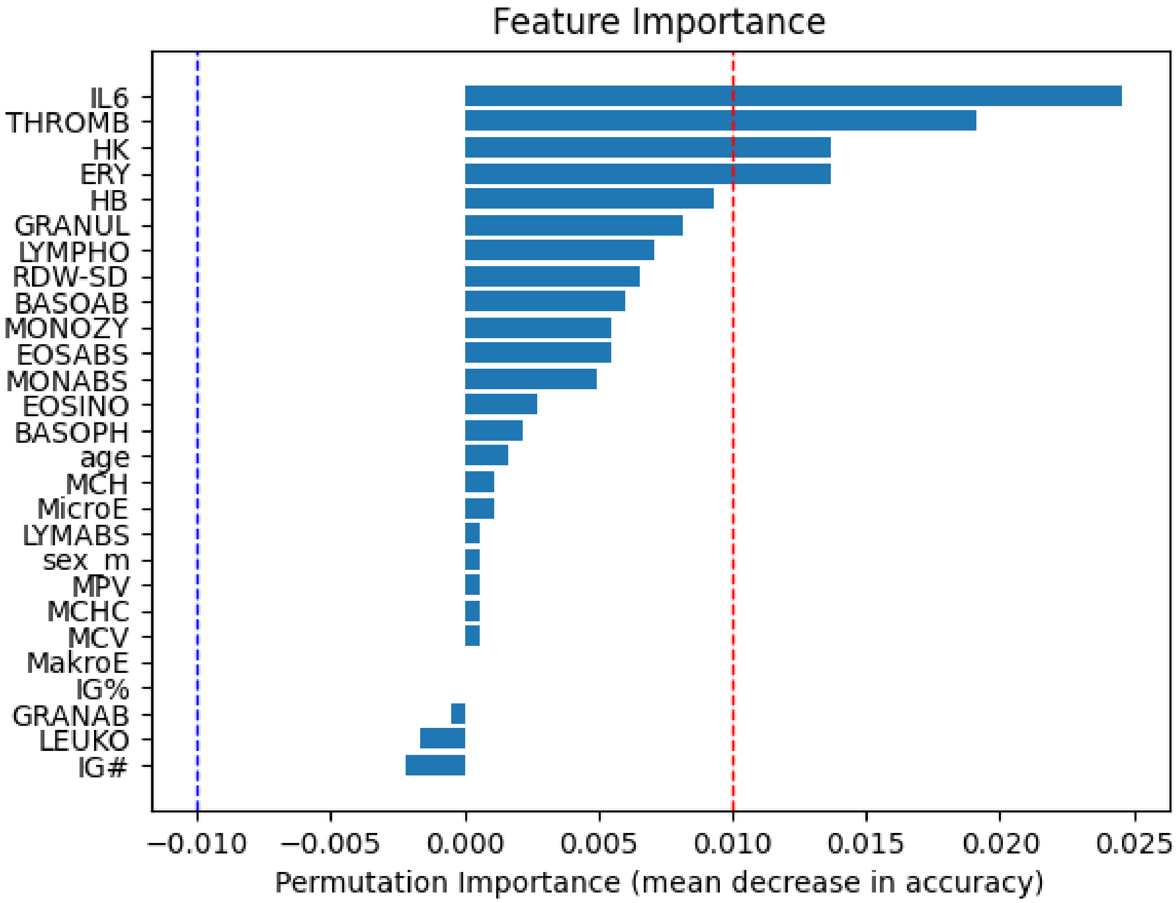


Figure 3.13: Permutation Importance with Neural Network

This threshold represents a quantitatively meaningful decrease in model accuracy. Features that surpass this threshold are Interleukin 6, Thrombocytes Count, Hematocrit and Erythrocytes Count. When shuffled, they lead to a more than 1% decrease in model accuracy, indicating their substantial role in the model's predictive ability. Also, from a practical standpoint, a 1% decrease in accuracy can be considered significant in clinical applications [12], where even a small decrease in predictive accuracy can substantially affect diagnostic decisions and patient outcomes.

6 Result of the Selection of Key Features

To consolidate the importance of the features, we employed a voting system that integrates the results from all the methods. This voting system allowed us to aggregate the findings, ensuring that the selected features are not only important in one model but consistently significant across multiple techniques. The system effectively highlighted features that repeatedly demonstrated relevance. Each method "voted" for a feature by assigning it either a 1 or 0, where a vote of 1 indicates that the method deemed the feature important, and a vote of 0 means the feature was not considered significant by that particular method. This binary voting process provided a straightforward way to quantify the importance of each feature.

The voting results, as depicted in Table 3.2, provide a clear summary of how each feature performed across the methods used.

Table 3.2: Summary of votes for each feature across six techniques

Feature	Box Plot	Pearson's Corr.	Spearman's Corr.	Random Forest	XGBoost	DL + PI	Total Votes
IL6	0	1	1	1	1	1	5
HK	1	0	1	1	1	1	5
LYMPHO	1	1	1	1	1	0	5
THROMB	0	0	1	1	1	1	4
HB	1	0	1	1	1	0	4
LYMABS	1	0	1	1	0	0	3
RDW-SD	0	0	1	1	1	0	3
ERY	1	0	1	1	0	0	3
EOSABS	1	0	1	1	0	0	3
BASOAB	1	0	1	1	0	0	3
Age	0	0	1	0	1	0	2
MONOZY	1	1	0	0	0	0	2
MPV	0	0	0	1	1	0	2
LEUKO	1	0	1	0	0	0	2
MakroE	0	0	0	0	1	0	1
EOSINO	0	0	0	0	1	0	1
BASOPH	1	0	0	0	0	0	1
Sex	0	0	0	0	0	0	0
GRANAB	0	0	0	0	0	0	0
MicroE	0	0	0	0	0	0	0
MCHC	0	0	0	0	0	0	0
IG%	0	0	0	0	0	0	0
IG#	0	0	0	0	0	0	0
GRANUL	0	0	0	0	0	0	0
MCV	0	0	0	0	0	0	0
MONABS	0	0	0	0	0	0	0

The criterion for feature retention was based on features that received a minimum of three out of six possible votes. This threshold was carefully chosen to ensure that only the most consistently recognized features were included in the final model development, based on several key considerations:

- **Statistical Consensus and Robustness:** Setting the threshold at half of the total votes emphasizes features that demonstrate consistency across various methods, mitigating the risk of including features relevant only under specific model configurations.
- **Mitigating the Dimensionality Problem:** This approach helps to reduce the dimensionality of the model, focusing on a smaller set of more impactful features, thereby improving the efficiency and performance of the model and avoiding the curse of dimensionality.

The following features were retained based on their vote counts:

- **Hematocrit and Interleukin 6**: Both features received 5 votes, consistently identified across all methods as crucial for the model's predictive capability.
- **Thrombocytes Count, Hemoglobin, and Erythrocytes**: These features received 4 votes each, emphasizing their strong relevance to sepsis prediction.
- **Basophils Percentage, Leukocytes Count, Eosinophils Percentage, Red Cell Distribution Width-Standard Deviation, and Lymphocytes Absolute Count**: Each of these features received 3 votes, indicating their importance but with slightly less consensus across the methodologies.

Conclusion

In conclusion, this chapter meticulously delineates the process of key feature selection using a variety of analytical techniques. This comprehensive approach highlights critical variables, emphasizing their significant contributions to predicting sepsis outcomes. The integration of multiple evaluation methods through a voting system reinforces the importance of these features, ensuring the final model is robust and capable of generalizing across diverse clinical settings. This rigorous selection process sets the stage for the subsequent phases of model implementation, crucial for developing a reliable predictive tool.

4

Model Implementation and Results

Introduction

In this chapter, we focus on the implementation of a machine learning model designed to predict sepsis at its early stages. Building upon the feature selection steps outlined in previous chapters, the primary objective here is to apply the selected model to the clinical datasets and evaluate their performance. The chapter also highlights the importance of hyperparameter tuning in improving model performance and reducing false positives and false negatives. A comparative analysis of the datasets is provided to determine the most effective approach for early sepsis prediction, setting the stage for potential clinical applications.

1 Sub-Datasets Construction

To enhance the robustness and relevance of our sepsis prediction models, we constructed two distinct subdatasets, each designed to test different aspects of predictive performance. These datasets incorporate varying levels of detail and complexity to provide a comprehensive evaluation of the features' predictive power.

1. The first subdataset is based on features from a benchmark dataset discussed in the literature review [64] conducted by Leipzig University, allowing us to replicate and validate their findings.
2. The second subdataset integrates both the reference study features and additional important features identified through our comprehensive feature selection process.

Table 4.1 provides a concise overview of the features included in each dataset.

Table 4.1: Overview of Dataset Features

Features	**First Dataset**	**Second Dataset**
Age	✓	✓
BASOAB		✓
BASOPH		
EOSABS		✓
EOSINO		
ERY	✓	✓
GRANAB		
GRANUL		
HB	✓	✓
HK		✓
IG#		
IG%		
IL6		✓
LEUKO	✓	✓
LYMABS		✓
LYMPHO		✓
MCV	✓	✓
MONABS		
MONOZY		
RDW-SD		✓
Sex	✓	✓
THROMB	✓	✓

2 Data Preprocessing

In this section, we delve into the crucial process of data preprocessing, which forms the foundation for achieving accurate and reliable results in machine learning. The importance of this step cannot be overstated, as it directly influences the quality of data and, consequently, the integrity of any conclusions drawn from it. The raw dataset presented for analysis comprised 109,897 records and 34 features, indicating a substantial volume of data to be refined.

To provide a clear visualization of the preprocessing workflow and its impact on the dataset, a comprehensive graph will be included. Figure 4.1 will detail the entire preprocessing process, with each step meticulously documented to offer a visual representation of how the raw data was transformed into a refined format ready for subsequent analysis.

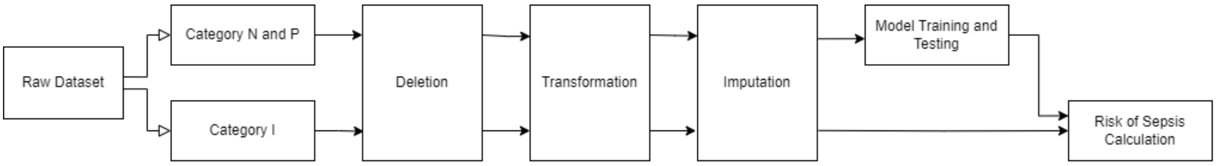


Figure 4.1: Data Preprocessing Diagram

2.1 Deletion

- **Elimination of Non-Sensible Categories:** We segmented our dataset based on the classification of sepsis status. While the 'Indeterminate' (Category: I) cases pose diagnostic ambiguities due to their uncertain sepsis status, they were not discarded. Instead, these cases were isolated and treated separately. As these indeterminate cases do not yield clear insights into whether a patient is sepsis-positive or sepsis-negative. Thus, for the purposes of training and testing our predictive model, we focused on a dataset comprising only the 'Sepsis-Negative' (Category: N) and 'Sepsis-Positive' (Category: P) groups. Simultaneously, the separated 'Indeterminate' dataset is reserved for making predictions, acknowledging the real-world scenario where the sepsis status of many patients may remain unclear at the point of care.
- **Removal of Irrelevant Columns:** During our preprocessing phase, a thorough review of the dataset's feature set revealed the presence of columns that were either irrelevant to the study's objectives or improperly formatted, representing potential sources of noise

and error in the subsequent analysis. After the removal of the complete typographical errors columns, the dataset's structure was refined to 30 columns.

- **Elimination of Non-Sensible Values:** Initially, our dataset included instances where missing data was represented by the value (-1) across various features. To quantify and address these missing values, we first calculated the count of '-1' entries for each feature. Upon identifying the features with missing data, we set a threshold to determine the feasibility of retaining these features. Specifically, we decided to drop any feature where the percentage of missing values exceeded 50%. This decision was based on the premise that a feature missing more than half of its data likely offers limited utility in predictive modeling and could potentially bias the results if inadequately addressed. Figure 4.2 illustrates the percentage of missing data across each feature within the dataset. A horizontal line at the 50% threshold clearly demarcates the limit set for acceptable levels of missingness. Features such as IL6 and IG-C are highlighted as they exceed this threshold and have therefore been excluded from further analysis.

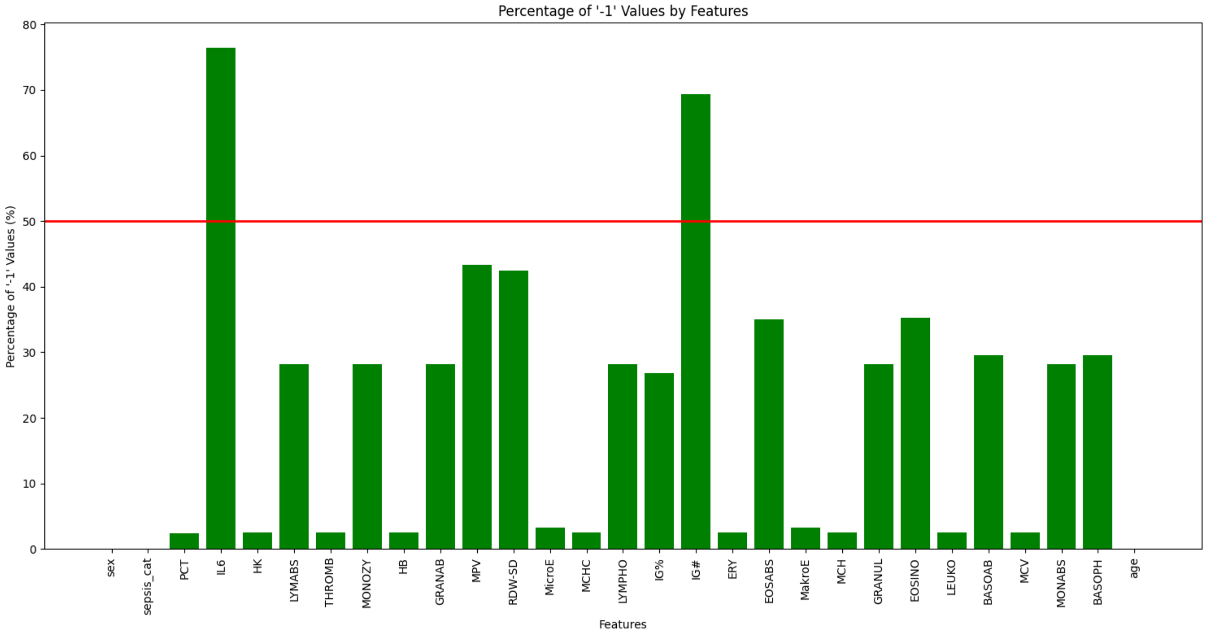


Figure 4.2: Percentage of Missing Features Across the Dataset

- **Specific Feature Attention:** Given the critical role of PCT levels in diagnosing sepsis, special attention was given to the handling of missing data in the 'PCT' feature. In our dataset, any instance where PCT data was absent, indicated by a (-1), was considered detrimental to the reliability and accuracy of our outcomes. Therefore, rows with missing PCT values were eliminated.

2.2 Transformation

- **Correction of Sex Labels and One Hot Encoding:** During the initial data exploration, we identified several records with non-standard sex labels that did not conform to the universally recognized categories of 'M' (Male) and 'W' (Female). These non-standard labels included typographical errors and ambiguous entries. To address this issue, we implemented a data cleaning step where all non-standard and ambiguous sex labels were identified and removed from the dataset.
- **Transformation of Birth Year to Age:** In our dataset, transitioning from recording birth years to directly noting age significantly enhances the dataset's utility for analytical processes.

2.3 Imputation

A critical component of preparing our dataset for robust analytical tasks involves addressing any remaining missing values post-initial preprocessing steps. To this end, we employed the KNN algorithm, a powerful method for imputing missing data based on the similarity of observations. KNN Imputation leverages the proximity of samples in the feature space to estimate missing values, assuming that points that are close to one another are likely to have similar data points. The rationale behind applying KNN Imputation post-feature selection is to ensure that only relevant features influence the imputation process. By selecting significant features first, we mitigate the risk of non-informative, noisy, or redundant features skewing the imputation of missing values.

3 Machine Learning Model Development

3.1 Adaptive Boosting

AdaBoost, short for Adaptive Boosting, is a seminal ensemble technique [46] designed to enhance the accuracy of weak classifiers through a sequential learning process where each subsequent classifier is tweaked to correct its predecessor's mistakes. Initially, all training instances are given equal weights, but as the algorithm progresses, AdaBoost focuses on instances that were previously misclassified by increasing their weights. This method not only amplifies the focus on harder-to-classify examples but also integrates the predictions of multiple weak learners to form a stronger final classifier. This strategy is particularly adept at refining the classifier's performance iteratively, making it highly effective for complex classification tasks where simple models (weak learners) fail to achieve good performance individually.

3.2 Random Under-Sampling

Random Under-Sampling (RUS) [26] tackles the issue of class imbalance by randomly removing samples from the over-represented class. This method is straightforward and helps to mitigate the bias towards the majority class commonly seen in imbalanced datasets. By equilibrating the class distribution, RUS simplifies the learning process, potentially reducing the time and computational resources needed for model training. However, the indiscriminate removal of data points can lead to the loss of critical information, which might affect the model's ability to generalize well to new data. This makes RUS inherently risky, especially in scenarios where every data instance might carry unique information crucial to the learning process.

3.3 RUSBoost Integration

3.3.1 Presentation of RUSBoost

The Random Under-Sampling Boosting (RUSBoost) algorithm [27] is an integration of the Random Under-Sampling (RUS) technique with the adaptive boosting method: AdaBoost. This hybrid approach is particularly designed to address class imbalance, a prevalent issue in datasets where one class significantly outnumbers the other. RUSBoost counteracts this imbalance by randomly eliminating instances from the majority class during each boosting iteration, thus aligning the class distribution more evenly and allowing the model to focus on the minority class, which often contains the critical cases in sepsis prediction. Our datasets exhibits a significant class imbalance, with non-sepsis cases vastly outnumbering sepsis instances. RUSBoost's under-sampling strategy effectively balances the class distribution, ensuring that the sepsis cases receive adequate attention during model training.

At the core of the RUSBoost architecture lies a sequence of weak learners (decision trees). Each learner is trained on a distinct version of the dataset that has been undersampled differently, ensuring diverse learning contexts and reducing the model's bias towards the majority class. The algorithm employs a weighting mechanism where misclassified instances in each iteration are assigned higher weights, thereby emphasizing the more difficult cases in subsequent iterations. This adaptive nature allows RUSBoost to continuously refine its focus on the samples that are most often misclassified, leading to a robust generalization performance.

3.3.2 RUSBoost Algorithm

All formulas related to the algorithm are sourced from the research paper [60]: Given a dataset $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $x_i \in X$ represents the feature vector and $y_i \in Y$ the corresponding class label for each of the m examples, RUSBoost enhances the representation of the minority class by selectively undersampling the majority class during the training process.

1. **Initialization:** Initially, each example (x_i, y_i) is assigned an equal weight:

$$D_1(i) = \frac{1}{m} \text{ [60]}$$

2. **Boosting Rounds:** The algorithm performs T boosting rounds, where each round t involves the following steps:

   (a) **Random Undersampling (RUS):** The dataset S is modified by randomly discarding examples from the majority class to achieve a desired class ratio $N\%$ of the minority class in the new training dataset S'_t.

   (b) **Training the Weak Learner:** The modified dataset S'_t and its corresponding weights D_t are used to train a weak learner, producing a hypothesis h_t.

   (c) **Calculating Pseudo-loss:** The pseudo-loss ϵ_t for the hypothesis h_t is calculated over the original dataset S, based on the current weights:

$$\epsilon_t = \sum_{i=1}^{m} D_t(i) \cdot \mathbf{1}(h_t(x_i) \neq y_i) \text{ [60]}$$

   where $\mathbf{1}$ is the indicator function that is 1 if $h_t(x_i) \neq y_i$ and 0 otherwise.

(d) **Updating the Weight Parameter:** The weight update parameter α_t is calculated as follows:

$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t} \text{ [60]}$$

(e) **Updating Weights:** The weights for the next iteration D_{t+1} are updated such that:

$$D_{t+1}(i) = \frac{D_t(i) \cdot e^{-\alpha_t \cdot \mathbf{1}(h_t(x_i)=y_i)}}{Z_t} \text{ [60]}$$

where Z_t is a normalization factor to ensure that D_{t+1} forms a valid probability distribution.

3. **Final Hypothesis:** After T iterations, the final hypothesis $H(x)$ is computed as a weighted sum of the weak hypotheses:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right) \text{ [60]}$$

3.4 Comparaison of AdaBoost, RUS and RUSBoost Integration

In the face of such challenges, the integration of RUSBoost presents a compelling advantage, particularly when compared to using AdaBoost and RUS independently. The paper by Seiffert et al. [60] explores this by integrating the simplicity and speed of RUS with the adaptive learning capability of AdaBoost, thus forming RUSBoost, which addresses class imbalance effectively by sequentially focusing more on the misclassified minority examples. the Table 4.2 based on the research [60] compares the advantages and disadvantages of AdaBoost, RUS and RUSBoost, highlighting how each approach addresses the class imbalance problem.

In our research, we chose RUSBoost primarily for its superior performance across various datasets, as demonstrated in empirical studies. This model outperforms both AdaBoost and RUS in handling imbalanced datasets by combining the advantages of boosting and undersampling without their individual drawbacks. RUSBoost consistently shows better classification accuracy and maintains robust performance metrics compared to AdaBoost and RUS alone, particularly in scenarios with severe class imbalances.

Model	Advantages	Disadvantages
AdaBoost	- Iteratively focuses on misclassified instances, improving classification accuracy. - Adaptive to changes in the learning process, adjusts weights to minimize errors. - Widely applicable and effective in many scenarios.	- Does not inherently address class imbalances, may bias towards the majority class. - Can be sensitive to noisy data and outliers. - Computationally intensive, especially with large datasets.
RUS	- Simple and fast to implement. - Reduces the size of the training data, speeding up the training process. - Effective in reducing the training time without complex computations.	- Random removal of majority class examples can lead to loss of important information. - Can cause underfitting, reducing the generalization ability of the model. - Does not adaptively focus on harder to classify instances.
RUSBoost	- Combines the strengths of AdaBoost and RUS, focusing on difficult cases without losing majority class information. - Mitigates the loss of information typical in RUS while benefiting from faster training times. - Generally performs better than using AdaBoost or RUS alone in imbalanced datasets.	- Still requires careful tuning of parameters like the number of boosting iterations and sampling level. - More complex to understand and implement compared to using RUS alone. - Can be computationally more intensive than simple RUS due to the boosting process.

Table 4.2: Comparison of Models with Advantages and Disadvantages

3.5 Hyperparameter Tuning Process

3.5.1 Selection of Hyperparameters

The selection of hyperparameters in the RUSBoost model is driven by the goal of maximizing the diagnostic accuracy while managing the risk of overfitting. The key hyperparameters adjusted during the tuning process include:

- **Estimator Max Depth:** The maximum depth of the individual decision trees (base estimators) is a critical parameter. Deeper trees can model more complex patterns but are also prone to overfitting. In this study, depths of 5, 10, and 15 were tested to determine the optimal complexity that balances bias and variance effectively.
- **Number of Estimators:** This parameter defines the number of sequential trees to

be constructed. It is directly proportional to the robustness of the model, with higher numbers typically improving model performance at the cost of increased computational demand and risk of overfitting. The values tested were 100, 300, 500, 700, and 900, providing a broad spectrum to assess the trade-offs between performance and efficiency.

- **Learning Rate:** The learning rate controls the contribution of each tree to the final model. A lower learning rate can often lead to better generalization capabilities at the expense of requiring more estimators to converge. Learning rates of 0.01, 0.1, 0.5, and 1 were explored to find an optimal setting that ensures rapid convergence without sacrificing accuracy.

3.5.2 Cross-Validation Strategy

Cross-validation is a vital step in the hyperparameter tuning process, especially when dealing with imbalanced datasets like ours. It ensures that the model is robust and generalizable by testing its performance on different subsets of the data.

Cross-validation involves splitting the dataset into multiple subsets or "folds." In each iteration, one fold is used as the test set, while the remaining folds are used for training the model. This process is repeated until each fold has served as the test set, and the final model performance is averaged across all folds.

In our study, we used stratified 5-fold cross-validation, a variant of cross-validation designed to address class imbalance. In stratified cross-validation, each fold maintains the original class distribution, ensuring that both the minority (sepsis) and majority (non-sepsis) classes are proportionately represented in every fold. This prevents biased training or evaluation scenarios and provides a more accurate measure of model performance on imbalanced data.

3.5.3 Hyperparameter Optimization with GridSearchCV

`GridSearchCV` was configured to:

- Perform a stratified 5-fold cross-validation for each combination of hyperparameters.
- Use ROC-AUC as the scoring metric to prioritize the model's ability to distinguish between positive (sepsis) and negative (non-sepsis) classes. The ROC-AUC score is

particularly suitable for our imbalanced datasets because it evaluates the model's performance across various threshold levels without being influenced by class distribution.

During the grid search process, each combination of hyperparameters was evaluated across all five folds. The mean ROC-AUC score from all folds was used to identify the best-performing combination. This strategy ensures that the selected hyperparameters not only perform well on a single fold but are robust and generalize effectively across multiple data splits.

By leveraging `GridSearchCV`, we systematically explored the hyperparameter space, selecting the configuration that maximized the ROC-AUC score. This method provided valuable insights into the model's stability and its ability to handle the complexities of imbalanced datasets.

3.5.4 Optimal Hyperparameters Across Datasets

After conducting a comprehensive hyperparameter tuning process using GridSearchCV coupled with stratified 5-fold cross-validation, the optimal set of parameters was identified for the RUSBoost model applied across the three distinct datasets. A total of 300 fits were executed, reflecting the fact that each of the 60 different configurations—across three max depths (5, 10, 15), five numbers of estimators (100, 300, 500, 700, 900), and four learning rates (0.01, 0.1, 0.5, 1)—was assessed five times due to the 5-fold cross-validation. The results consistently highlighted the same configuration as most effective, reflecting the robustness and generalizability of the chosen hyperparameters under varying conditions.

The best hyperparameters determined for the RUSBoost algorithm across all datasets are as follows:

- **Estimator Max Depth:** 5
- **Learning Rate:** 0.01
- **Number of Estimators:** 300

This configuration suggests that a relatively shallow decision tree depth, combined with a low learning rate and a moderate number of trees, provides the best balance between bias and variance, thus enhancing the model's ability to generalize well to unseen data. The choice of a shallow tree depth helps in preventing the model from overfitting, particularly crucial in medical diagnostic applications where the cost of false predictions can be high. Moreover, the low learning rate of 0.01 ensures that each new tree makes a conservative contribution to the overall model, allowing for a gradual, but more robust learning process. Lastly, the selection of 300 trees strikes an optimal balance between computational efficiency and predictive performance, making it a practical choice for real-world applications.

4 Machine Learning Model Results

4.1 Results of Performance Metrics

The evaluation of the RUSBoost algorithm across the different datasets yielded consistent and promising results, particularly in terms of ROC AUC and recall, which are critical metrics for the early prediction of sepsis. Table 4.3 presents the best performing metrics.

Table 4.3: Performance Metrics of the Five-Fold Cross-Validation

Metric	Dataset 1	Dataset 2
ROC AUC	0.9167	**0.9307**
Accuracy	0.8507	**0.8652**
Precision	0.5257	**0.5577**
Recall	0.8256	**0.8445**
F1-Score	0.6407	**0.6697**

The ROC AUC and Recall metrics were highlighted as the most critical for this study due to their relevance in medical diagnostic performance:

ROC AUC: High values across all datasets (0.9167, 0.9307) suggest that the model possesses excellent discriminatory power. This is vital in clinical settings where the early and accurate detection of sepsis can significantly influence treatment outcomes.

Recall: High recall values (0.8256,0.8445) indicate that the model successfully identifies a high percentage of actual sepsis cases. In the context of sepsis, where failing to detect true positives (patients with sepsis) could lead to severe complications or death, the importance of recall cannot be overstated.

The consistent performance across different datasets verifies the robustness of the RUSBoost model configured with the optimized hyperparameters. The high scores in

ROC AUC and recall across all datasets underscore the model's capability to serve as a reliable tool in the early prediction of sepsis, aligning with the primary research objectives.

The precision, while lower compared to recall, reflects the challenging nature of predicting sepsis, where the consequences of missing a true positive are more critical than the cost of false positives. The balance between precision and recall is captured by the F1-score, which remains reasonably high across datasets : (0.6407 , 0.6697) indicating a moderately balanced model.

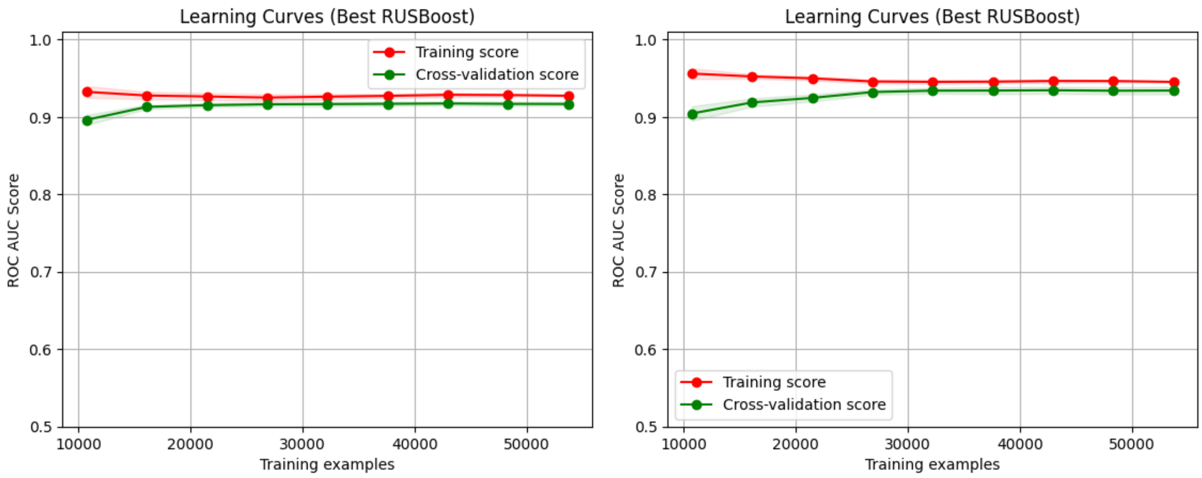
These results collectively demonstrate that the model can effectively assist healthcare professionals by providing timely and accurate predictions of sepsis, potentially improving patient outcomes through earlier intervention and targeted treatment.

4.2 Model Evaluation

A learning curve is a graphical representation that shows the progress of the model's learning over time, measured by the change in performance, such as the ROC AUC score, against the volume of training examples used. Each curve in the figure depicts this relationship and is crucial for assessing the model's tendency to overfit or underfit the data.

The learning curves of Figure 4.3 generated for the RUSBoost model across the datasets provide valuable insights into the model's performance dynamics over the training process. Each curve represents the change in ROC AUC score as the number of training examples increases.

Notably, both learning curves converge as the number of training examples increases, indicating that the model is stabilizing and achieving a balance between learning from the training data and generalizing to new, unseen data. This convergence is a positive sign that the model is capable of consistent performance, avoiding high variance or high bias when exposed to varying datasets.



(a) Dataset 1

(b) Dataset 2

Figure 4.3: Learning curves of the RUSBoost Model Across Datasets

- **Training Score Trends:**
  - The training scores start high and remain relatively stable as the number of training examples increases.
  - This stability in training scores, consistently close to an ROC AUC score of 0.9 or above, suggests that the model is sufficiently complex to capture the underlying patterns in the data without being too simplistic.
  - There is no apparent degradation in performance, indicating no underfitting.
- **Validation Score Trends:**
  - The cross-validation scores begin slightly lower than the training scores but quickly converge to a similar level as more training data is used.
  - The narrowing gap between the training and validation scores as more data points are added indicates that the model generalizes well.
  - The validation curves leveling close to the training curves, without significant fluctuations, are indicative of good model performance with a balanced variance and bias.
- **Consistency Across Datasets:** The consistency of the learning curves across multiple datasets underscores the model's robustness and its ability to generalize across different scenarios without the need for extensive re-tuning or modification.
- **Stable Performance with Increased Data:** The model's performance does not degrade with the addition of more data, which often challenges other models due to

overfitting or increased noise. Instead, the RUSBoost demonstrates enhanced stability and reliability, suggesting that it can handle larger and potentially more complex datasets without a loss in performance quality.

Overall, these learning curves provide compelling evidence that the RUSBoost model is well-tuned for the task of sepsis prediction, offering high reliability and excellent potential for deployment in practical settings where new, unseen patient data are regularly encountered.

4.3 Best Performing Dataset and Benchmarking

In our study, we benchmarked our sepsis prediction models against existing clinical baselines, with a particular focus on a recent study that applied machine learning techniques to predict sepsis using CBC [64] of non-ICU data from a german tertiary carecenter. Our model utilized a dataset from the Klinikum of Passau, featuring the same demographic and CBC parameters. This provided a direct comparison point to assess the relative performance of similar models across different datasets. Table 4.4 provides a comparative analysis between the baseline study and our own, across these two distinct datasets.

Table 4.4: Comparative Analysis of Sepsis Prediction Datasets Results

Metrics / Study	**Reference Study**	**Our Study - Dataset 1**	**Our Study - Dataset 2**
Dataset	Demographics and Basic CBC Parameters	Demographics and Basic CBC Parameters	Demographics and Basic CBC Parameters + Important Key features
AUROC score	0.805 - 0.872	0.9167	0.9307
Accuracy	N/A	85.07%	86.52%
Precision	N/A	52.57%	55.77%
Recall	N/A	82.56%	84.45%
F1-Score	N/A	64.07%	66.97%
External Validation	0.805 (Greifswald), 0.845 (MIMIC-IV)	N/A	N/A

The referenced study achieved notable predictive accuracy with an Area Under the Receiver Operating Characteristic (AUROC) scores ranging from 0.805 to 0.872 across various validation scenarios. In our initial comparisons using the Klinikum of Passau dataset, our model demonstrated comparable performance, achieving an AUROC score of 0.9167, which indicates robust diagnostic capabilities in a similar non-ICU setting.

Furthermore, to enhance the predictive power of our model, we incorporated additional clinical features deemed relevant based on recent medical insights and our initial model assessments. These features included: HK, BASOPH, EOSINO, RDW-SD, LYMPHO and LYMABS. By integrating these features, our extended model, applied to a second dataset, showed improved performance, with an increased AUROC of 0.9307, thereby suggesting that the additional parameters could provide significant benefits in clinical settings for early sepsis detection.

4.4 Feature Importance

The importance of a feature is measured based on how effectively the feature contributes to the model's ability to reduce uncertainty or impurity in the data during the training process.

The bar chart of the Figure 4.4 illustrates the relative importance of each feature in predicting sepsis.

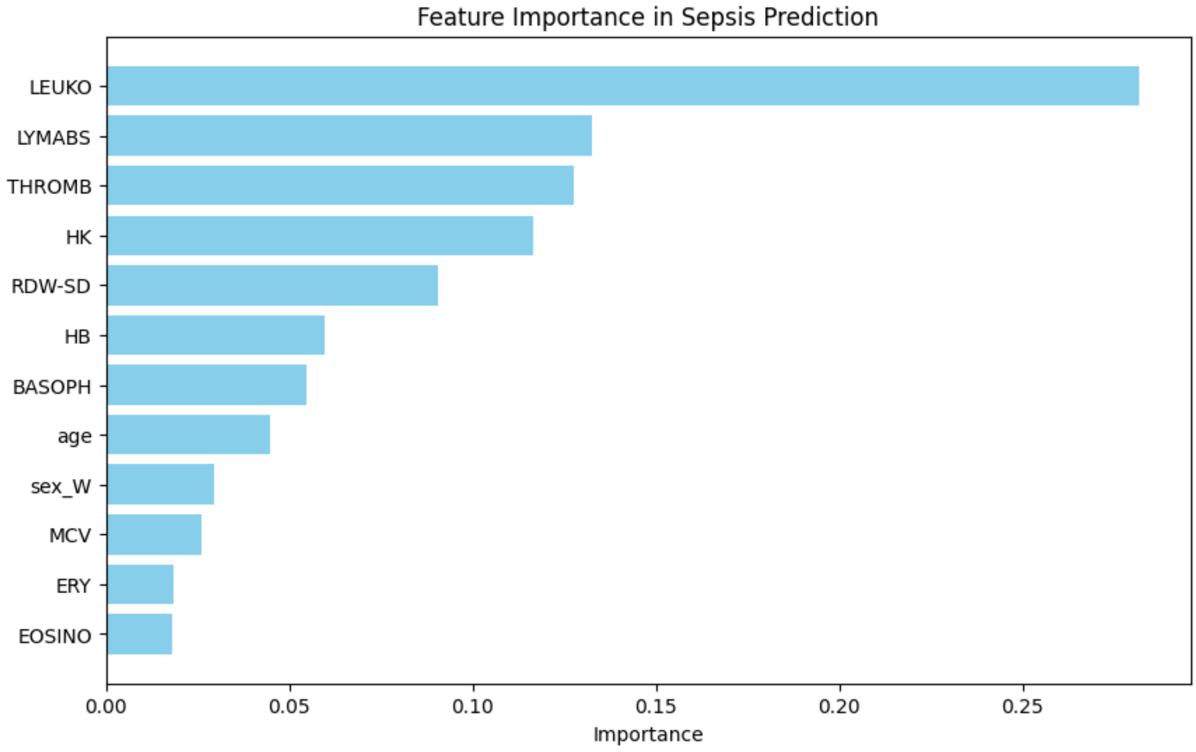


Figure 4.4: Feature Importance of RUSBoost Classifier

Notably, LEUKO, emerged as the top most significant predictor. This finding aligns with clinical expectations, as abnormalities in the white blood cells are commonly associated with infection and systemic inflammation, hallmark characteristics of sepsis. LYMABS, THROMB and HK, another critical features, along with RDW-SD, underscores the role of white and red blood cell metrics in sepsis diagnostics. This insight is particularly useful for clinicians, as these measures provide indirect clues about the presence of anemia or other hematological dysfunctions associated with sepsis.

4.5 Permutation Importance

To further refine our understanding of feature contributions towards predicting sepsis, we employed the permutation importance technique. This method involves systematically shuffling each feature in the test set and measuring the impact on the model's performance across various metrics. The results were compiled into the Table 4.5, summarizing the mean impact of shuffling each feature on the model's performance metrics.

Table 4.5: Impact of Different Features on Model Metrics

Feature/Impact on:	Accuracy	AUROC	Precision	Recall	F1 Score
LEUKO	0.058453	0.066692	0.117068	0.122998	0.124592
THROMB	0.041339	0.048697	0.085404	0.130074	0.103117
LYMABS	0.014990	0.035416	0.028390	0.092318	0.050428
HK	0.005247	0.014995	0.008589	0.043389	0.020195
age	0.003854	0.004351	0.008100	0.010289	0.009132
sex	0.003131	0.002568	0.006733	0.006611	0.006977
MCV	0.001878	0.000981	0.004244	0.001723	0.003626
BASOPH	0.001081	0.003232	0.000722	0.019600	0.006814
EOSINO	0.000775	0.000598	0.001575	0.002700	0.001997
RDW-SD	-0.000976	0.007453	-0.005766	0.033799	0.006948
ERY	-0.002445	-0.000145	-0.006575	0.007263	-0.002397
HB	-0.002818	0.002178	-0.008307	0.014944	-0.001114

4.6 Confusion Matrix in Clinical Context

4.6.1 Detailed Confusion Matrix Analysis

The confusion matrix is a critical evaluation tool in classification problems, particularly in medical settings where the consequences of different types of errors vary significantly. For our sepsis prediction model, the confusion matrix provides a detailed breakdown of the model's performance in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Below, we present the aggregated confusion matrix from a five-fold cross-validation on the dataset:

	Predicted Non-Sepsis	**Predicted Sepsis**
Actual Non-Sepsis	48201	8139
Actual Sepsis	1873	8867

Table 4.6: Aggregated Confusion Matrix

- TP: 8867 cases where the model correctly identified patients with sepsis.
- TN: 48201 cases where the model correctly identified patients without sepsis.
- FP: 8139 cases where the model incorrectly predicted sepsis in patients without it.
- FN: 1873 cases where the model failed to detect sepsis in patients who actually had it.

4.6.2 Real-world Implications

The real-world implications of these results are significant:

- **Correct Diagnoses:** The high number of true positives and true negatives suggests that the model is capable of correctly identifying a majority of sepsis and non-sepsis cases, potentially allowing for timely and appropriate treatment interventions.
- **False Alarms:** The false positives, while undesirable, are less detrimental in this context than false negatives. False alarms may lead to unnecessary treatments or additional diagnostic tests, which, while costly and potentially stressful for patients, are less likely to result in severe harm.
- **Missed Sepsis Cases:** The false negatives are particularly concerning in the context of sepsis, where delayed treatment can significantly worsen patient outcomes, including increased mortality rates. Reducing false negatives is crucial, even if it comes at the cost of an increase in false positives.

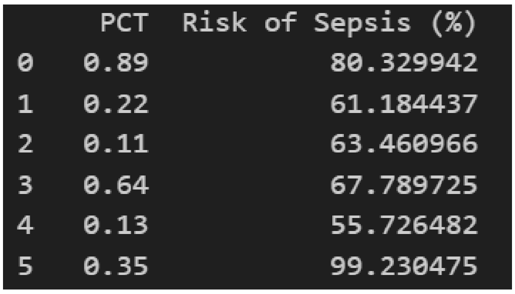
4.6.3 Cost-Benefit Analysis

In evaluating the trade-offs between different types of errors, the model's performance can be seen in light of the following:

- **Cost of False Positives:** While false positives can lead to unnecessary medical interventions, the costs associated here include additional tests, increased healthcare resource usage, and the psychological impact on patients and their families. However, these costs must be weighed against the potential life-saving benefits of ensuring no cases of sepsis are missed.
- **Cost of False Negatives:** The cost of false negatives is generally more severe, as a missed diagnosis of sepsis can lead to delayed treatment, resulting in rapid deterioration of the patient's condition, increased likelihood of severe complications, extended hospital stays, and even increased mortality rates. The financial costs here are also significant, stemming from prolonged intensive care and increased resource utilization, but the human cost in terms of patient outcomes is far greater. The costs here are not only medical but also involve significant ethical considerations. Preventing these cases should be a priority, even if it means accepting a higher number of false positives.

4.7 Case Studies and Examples

We present the outcomes of this model on the category 'I' that was initially eliminated due to the indeterminate state of sepsis, emphasizing the calculated probabilities of sepsis which demonstrate the model's operational capabilities. Each example of the output on Figure 4.5 represents a patient's predicted probability of sepsis against their actual PCT value



```
    PCT  Risk of Sepsis (%)
0  0.89           80.329942
1  0.22           61.184437
2  0.11           63.460966
3  0.64           67.789725
4  0.13           55.726482
5  0.35           99.230475
```

Figure 4.5: Model Output

5 Evaluation of the Sepsis Prediction Model

5.1 Positive Aspects of the Model

Improved Performance over Reference Models: Our model demonstrated superior performance compared to the benchmarks set by previous studies. This indicates not only the robustness of our predictive algorithms but also possibly the higher quality or more relevant nature of our dataset.

Enhanced Predictive Metrics: The focus on recall as a critical metric for the prediction of sepsis is particularly noteworthy. The model's ability to maintain a high recall ensures that fewer cases of sepsis go undetected, which is vital for timely medical intervention.

Comprehensive Feature Analysis: Through iterative feature selection processes, our model refined the number of features from 26 down to an optimal subset, enhancing model simplicity and focus while reducing overfitting potential. This approach helped in isolating the most impactful predictors such as LEUKO, which showed significant importance in permutation importance evaluations.

5.2 Challenges and Limitations

Indeterminate Category Testing: A major challenge was the inability to verify the true labels for predictions made in the 'Indeterminate' category (Category I). This limitation restricts our ability to assess the model's accuracy in these cases definitively. Future validations could benefit from external datasets where true sepsis labels are known, allowing for a more thorough accuracy assessment.

Potential for Overfitting with Multiple Features: Initially, a broad array of features was considered, but this was narrowed down significantly to manage model complexity and computational efficiency. However, there remains a risk that some of the discarded features, especially those detailing white blood cell counts beyond LEUKO, could provide valuable insights. A more nuanced exploration of related sub-features might enhance the model's predictive power and reduce the noise-to-signal ratio further.

Lack of External Validation: Currently, the model's validation relies heavily on internal datasets and cross-validation methods. External validation with independent datasets is crucial for verifying the model's generalizability and robustness in different clinical settings or geographical locations.

5.3 Recommendations for Future Research

External Dataset Testing: To overcome the limitations noted with Category I testing and enhance the credibility of the model, future studies should aim to test the model using external datasets with confirmed sepsis labels. This will help in establishing the model's effectiveness across different populations and clinical environments.

Expanded Feature Analysis: Further research should consider revisiting the discarded features and exploring a more granular level of blood markers, especially those related to immune response, to potentially uncover nuanced patterns that could improve prediction accuracy.

Integration with Clinical Practice: It would be beneficial to integrate the model's predictions with real-time clinical decision-making processes to evaluate its practical utility and impact on patient outcomes. Pilot studies in clinical settings could provide valuable feedback for refining the model.

Continual Learning and Update Mechanisms: Implementing mechanisms for the model to learn continually from new data and updating its parameters accordingly could help in maintaining its relevance and accuracy over time as new medical findings and technologies emerge.

Conclusion

The implementation of a machine learning model in this chapter demonstrated significant progress in improving the early detection of sepsis. RUSBoost showed high performance in handling imbalanced data, achieving high accuracy and reliable prediction outcomes. The results confirmed the potential of machine learning techniques in analyzing clinical datasets and extracting meaningful patterns that could be used for sepsis prediction.

Conclusion and Perspectives

This thesis investigated the application of machine learning algorithms in improving early sepsis detection. The findings confirm that machine learning can significantly enhance healthcare by reducing diagnosis times and enabling timely interventions. This cost-effective approach utilizes readily available clinical data, making it feasible for widespread implementation in hospitals.

Nevertheless, the study also identified limitations, such as the variability in patient data and the complexities involved in optimizing machine learning models. Future efforts must focus on refining these algorithms and validating models across diverse healthcare settings. Integrating these tools into clinical workflows remains crucial for future research and development.

Looking ahead, enhancing the generalizability of these models by testing on larger, diverse datasets from various healthcare facilities is essential. This will verify algorithm robustness and consistent performance across different settings.

Another future direction is to integrate these models with real-time hospital patient monitoring systems. By continuously analyzing data, these tools could offer continuous risk assessments and alert medical teams to emerging sepsis cases, shifting machine learning from a diagnostic to a proactive patient care tool.

In summary, this thesis lays a groundwork for advancing healthcare diagnostics with machine learning, which holds significant potential to transform sepsis detection and other healthcare challenges, leading to better, quicker, and more tailored patient care..

References

[1] Crisp-dm help. Accessed: 2024-10-02.

[2] Xgboost: A powerful ml model for classification and regression. https://machinelearningmodels.org/xgboost-a-powerful-ml-model-for-classification-and-regression/. Accessed: 2024-09-29.

[3] SMOB Abdullah, RH Sørensen, and FE Nielsen. Prognostic accuracy of sofa, qsofa, and sirs for mortality among emergency department patients with infections. *Infect Drug Resist*, 14:2763–2775, 2021.

[4] Luisa Agnello, Rosaria Vincenza Giglio, Giulia Bivona, Concetta Scazzone, Caterina Maria Gambino, Alessandro Iacona, Anna Maria Ciaccio, Bruna Lo Sasso, and Marcello Ciaccio. The value of a complete blood count (cbc) for sepsis diagnosis and prognosis. *Diagnostics*, 11:1881, 2021.

[5] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky. Epidemiology of severe sepsis in the united states: Analysis of incidence, outcome, and associated costs of care. *Crit Care Med*, 29(7):1303–10, 7 2001.

[6] Derek C. Angus and Tom van der Poll. Severe sepsis and septic shock. *New England Journal of Medicine*, 369(9):840–851, 2013.

[7] Eyal Bloch, Tomer Rotem, Jonathan Cohen, Philip Singer, and Yitzhak Aperstein. Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction. *Journal of Healthcare Engineering*, 2019:5930379, 2019.

[8] Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland M H Schein, and William J Sibbald. Definitions for sepsis

and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, 1 1992.

[9] J. E. Camacho-Cogollo, I. Bonet, B. Gil, and E. Iadanza. Machine learning models for early prediction of sepsis on large healthcare datasets. *Electronics*, 11(9), 5 2022.

[10] Javier Enrique Camacho-Cogollo, Isis Bonet, Bladimir Gil, and Ernesto Iadanza. Machine learning models for early prediction of sepsis on large healthcare datasets. *Electronics*, 11(9):1507, 2022.

[11] The Surviving Sepsis Campaign. Post-sepsis morbidity and mortality study. 2017.

[12] Carlos Castaneda, Kimberly Nalley, Conor Mannion, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, 5(4), 2015.

[13] A. Chandorkar and J. Simkins. Emerging fungal cutaneous infections in immunocompromised patients. *Current Fungal Infection Reports*, 14:217–224, 9 2020. Published 09 June 2020.

[14] Jian Chen and Haiming Wei. Immune intervention in sepsis. *Frontiers in Pharmacology*, 12, 2021.

[15] J. Chertoff, M. Chisum, B. Garcia, et al. Lactate kinetics in sepsis and septic shock: a review of the literature and rationale for further research. *Journal of Intensive Care*, 3:39, 2015.

[16] A. Creamer and J. Keep. Imaging in severe sepsis and septic shock: is early radiological identification of occult sources of infection needed? *Critical Care*, 18(Suppl 2):P12, 2014.

[17] G.P. Dobson, H.L. Letson, and J.L. Morris. Revolution in sepsis: a symptoms-based to a systems-based approach? *Journal of Biomedical Science*, 31, 2024. Received 02 April 2024; Accepted 17 May 2024; Published 30 May 2024.

[18] Siddharth Dugar, Chirag Choudhary, and Abhijit Duggal. Sepsis and septic shock: Guideline-based management. *Cleveland Clinic Journal of Medicine*, 87(1):53–64, 2020.

[19] C.F. Duncan, T. Youngstein, M.D. Kirrane, et al. Diagnostic challenges in sepsis. *Current Infectious Disease Reports*, 23(22), 2021. Accepted 24 September 2021; Published 25 October 2021.

[20] Birte Dyck, Matthias Unterberg, Michael Adamzik, and Björn Koos. The impact of pathogens on sepsis prevalence and outcome. *Pathogens*, 13(1), 2024.

[21] Find Lab Test. Procalcitonin in online lab tests stores. https://www.findlabtest.com/lab-test/general-wellness/procalcitonin-quest-16265, 2024. Accessed: September 2024.

[22] M. Fisher, L. Golestaneh, M. Allon, K. Abreo, and M.H. Mokrzycki. Prevention of bloodstream infections in patients undergoing hemodialysis. *Clin J Am Soc Nephrol*, 15(1):132–151, 2020. Epub 2019 Dec 5. Erratum in: Clin J Am Soc Nephrol. 2022 Apr;17(4):568-569. doi: 10.2215/CJN.01840222.

[23] C. Fleischmann, A. Scherag, N. K. Adhikari, C. S. Hartog, T. Tsaganos, P. Schlattmann, D. C. Angus, K. Reinhart, and International Forum of Acute Care Trialists. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. *Am J Respir Crit Care Med*, 193(3):259–72, 2 2016.

[24] Robert Gauer, Damon Forbes, and Nathan Boyer. Sepsis: Diagnosis and management. *American Family Physician*, 101(7):409–418, 2020.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2009.

[26] Imbalanced-learn. imblearn.under_sampling.randomundersampler. https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html. Accessed: 2024-10-02.

[27] imbalanced-learn developers. Rusboostclassifier — imbalanced-learn 0.8.0 documentation, 2024. Accessed: September 30, 2024.

[28] Khandaker Reajul Islam, Johayra Prithula, Jaya Kumar, Toh Leong Tan, Mamun Bin Ibne Reaz, Md. Shaheenur Islam Sumon, and Muhammad E. H. Chowdhury. Machine learning-based early prediction of sepsis using electronic health records: A systematic review. *Journal of Clinical Medicine*, 12(17), 2023.

[29] Hyun-Joo Kam and Hyun-Young Kim. Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine*, 89:248–255, 2017.

[30] Keras developers. Lstm layer — keras api documentation, 2024. Accessed: September 30, 2024.

[31] M.M.A. Kip, J.A. van Oers, A. Shajiei, et al. Cost-effectiveness of procalcitonin testing to guide antibiotic treatment duration in critically ill patients: results from a randomised controlled multicentre trial in the netherlands. *Critical Care*, 22(293), 2018.

[32] A. Kumar, D. Roberts, K.E. Wood, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med*, 34(6):1589–1596, 2006. Posted in Critical Care, ICU Trials, Landmark Trials on October 18, 2013.

[33] Mitchell M Levy, Mitchell P Fink, John C Marshall, Edward Abraham, Derek Angus, Deborah Cook, Jonathan Cohen, Steven M Opal, Jean-Louis Vincent, and Graham Ramsay. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Critical Care Medicine*, 31(4):1250–1256, 4 2003.

[34] E.C.N. Luijks, E.C. van der Slikke, A.R.H. van Zanten, et al. Societal costs of sepsis in the netherlands. *Critical Care*, 28:29, 2024.

[35] Amanda McCoy and Rajarshi Das. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Quality*, 6:e000158, 2017.

[36] R. S. Nannan Panday, E. M. J. Lammers, N. Alam, et al. An overview of positive cultures and clinical outcomes in septic patients: a sub-analysis of the prehospital antibiotics against sepsis (phantasi) trial. *Critical Care*, 23:182, 2019.

[37] University of Illinois. More machine learning methods: Random forests, 2023. Accessed: 2024-09-29.

[38] F. Pandolfi, C. Brun-Buisson, D. Guillemot, et al. Care pathways of sepsis survivors: sequelae, mortality and use of healthcare services in france, 2015–2018. *Critical Care*,

27, 2023. Received 18 September 2023; Accepted 08 November 2023; Published 10 November 2023.

[39] C. Pierrakos, D. Velissaris, M. Bisdorff, J.C. Marshall, and J.L. Vincent. Biomarkers of sepsis: time for a reappraisal. *Critical Care*, 24(1), 6 2020.

[40] ReadMedium. Xgboost feature importance: Gain vs weight & intuition behind it, 2020. Accessed: 2024-09-29.

[41] Chanu Rhee, Michael V. Murphy, Lingling Li, Richard Platt, Michael Klompas, for the Centers for Disease Control, and Prevention Epicenters Program. Comparison of trends in sepsis incidence and coding using administrative claims versus objective clinical data. *Clinical Infectious Diseases*, 60(1):88–95, 2014.

[42] A. Rhodes, L. E. Evans, W. Alhazzani, M. M. Levy, M. Antonelli, R. Ferrer, A. Kumar, J. E. Sevransky, C. L. Sprung, M. E. Nunnally, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. *Critical Care Medicine*, 45(3):486–552, 2017.

[43] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R Machado, Konrad K Reinhart, Kathryn Rowan, Christopher W Seymour, R Scott Watson, T Eoin West, Fatima Marinho, Simon I Hay, Rafael Lozano, Alan D Lopez, Derek C Angus, Christopher J L Murray, and Mohsen Naghavi. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020.

[44] Kristina E Rudd, Scott C Johnson, Kareha M Agesa, Katya A Shackelford, Derrick Tsoi, Daniel R Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020.

[45] T. Schupp, K. Weidner, J. Rusnak, et al. C-reactive protein and procalcitonin during course of sepsis and septic shock. *Irish Journal of Medical Science*, 193:457–468, 2024.

[46] Scikit-learn. sklearn.ensemble.adaboostclassifier. https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.AdaBoostClassifier.html. Accessed: 2024-10-02.

[47] scikit-learn developers. Adaboostclassifier — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[48] scikit-learn developers. f1_score — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[49] scikit-learn developers. Naive bayes — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[50] scikit-learn developers. precision_score — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[51] scikit-learn developers. Randomforestclassifier — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[52] scikit-learn developers. recall_score — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[53] scikit-learn developers. roc_auc_score — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[54] scikit-learn developers. Stackingclassifier — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[55] scikit-learn developers. Supervised neural networks — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[56] scikit-learn developers. Support vector machines — scikit-learn 1.0.2 documentation, 2021. Accessed: September 30, 2024.

[57] scikit-learn developers. accuracy_score — scikit-learn development version documentation, 2024. Accessed: September 30, 2024.

[58] scikit-learn developers. Kneighbors classifier — scikit-learn development version documentation, 2024. Accessed: September 30, 2024.

[59] scikit-learn developers. Logistic regression — scikit-learn 1.5 documentation, 2024. Accessed: September 30, 2024.

[60] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(1):185–197, Jan 2010.

[61] Sifa-Sibe. Hunderttausende vermeidbare todesfälle, 2020. Accessed: 2024-09-28.

[62] Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2016.

[63] Mervyn Singer, Clifford S Deutschman, Christopher W Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, Richard S Hotchkiss, Mitchell M Levy, John C Marshall, Greg S Martin, Steven M Opal, Gordon D Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2 2016.

[64] D. Steinbach, P. C. Ahrens, M. Schmidt, M. Federbusch, L. Heuft, C. Lübbert, M. Nauck, M. Gründling, B. Isermann, S. Gibb, and T. Kaiser. Applying machine learning to blood count data predicts sepsis with icu admission. *Clinical Chemistry*, 70(3):506–515, 3 2024.

[65] T2 Biosystems. Sepsis information page. Online, 2024. Accessed: 2024-09-28.

[66] Sudhir U, R K Venkatachalaiah, T A Kumar, M Y Rao, and P Kempegowda. Significance of serum procalcitonin in sepsis. *Indian J Crit Care Med*, 15(1):1–5, 2011.

[67] D. Wang, J. Li, Y. Sun, X. Ding, X. Zhang, S. Liu, B. Han, H. Wang, X. Duan, and T. Sun. A machine learning model for accurate prediction of sepsis in icu patients. *Frontiers in Public Health*, 9:754348, 2021.

[68] World Health Organization. Sepsis. https://www.who.int/news-room/fact-sheets/detail/sepsis, 5 2024.

[69] XGBoost developers. Xgboost documentation — stable version, 2024. Accessed: September 30, 2024.

[70] Zhenhua Yang, Xian Cui, and Zhi Song. Predicting sepsis onset in icu using machine learning models: a systematic review and meta-analysis. *BMC Infectious Diseases*, 23:635, 2023.

[71] H. A. Zaki, S. Bensliman, K. Bashir, et al. Accuracy of procalcitonin for diagnosing sepsis in adult patients admitted to the emergency department: a systematic review and meta-analysis. *Systematic Reviews*, 13:37, 2024.

REPUBLIC OF TUNISIA
Ministry of Higher Education and Scientific Research
University of Carthage

National Institute of Applied Sciences and Technology

التنبؤ المبكر بالإنتان باستخدام خوارزميات التعلم الآلي

تقدم هذه الأطروحة إطار عمل للتعلم الآلي لتعزيز الكشف المبكر عن الإنتان، مستفيدة من الإمكانات التنبؤية لمقاييس اختبارات الدم الروتينية ومستويات البروكالسيتونين (PCT). تستخدم الدراسة التحليلات البيانية المتقدمة لفحص العلاقات المتبادلة بين سلوكيات عد الدم والبروكالسيتونين، مع تحديد الأنماط الدقيقة التي تشير إلى بدء الإنتان. من خلال اختبار الخوارزميات الدقيق وتعديل المعايير، تهدف البحوث إلى تطوير أداة تشخيصية فعالة من حيث التكلفة وموثوقة تقلل بشكل كبير من وقت التشخيص. تعد هذه الطريقة بتحسين نتائج المرضى من خلال تمكين التدخلات الأكثر دقة وأبكر في علاج الإنتان.

الكلمات المفاتيح: التعلم الآلي، كشف الإنتان، التحليلات التنبؤية، مؤشرات الدم الحيوية، تعديل المعايير، دقة التشخيص، تحليل اختبارات الدم، تكنولوجيا الرعاية الصحية.

Prédiction précoce de la sepsis utilisant des algorithmes d'apprentissage automatique

Cette thèse présente un cadre d'apprentissage automatique pour améliorer la détection précoce de la sepsis, en exploitant le potentiel prédictif des mesures des tests sanguins de routine et des niveaux de Procalcitonine (PCT). L'étude utilise des analyses de données avancées pour examiner les interrelations entre les comportements des comptages sanguins et le PCT, identifiant des motifs subtils qui signalent le début de la sepsis. Grâce à des tests rigoureux d'algorithmes et au réglage des hyperparamètres, la recherche vise à développer un outil diagnostique fiable et rentable qui réduit considérablement le temps de diagnostic. Cette approche promet d'améliorer les résultats pour les patients en permettant des interventions plus précoces et plus précises dans le traitement de la sepsis.

Mots-clés : Apprentissage automatique, Détection de la sepsis, Analytique prédictive, Biomarqueurs sanguins, Réglage des hyperparamètres, Exactitude diagnostique, Analyse de tests sanguins, Technologie de la santé.

Early Prediction of Sepsis using Machine Learning Algorithms

This thesis presents a machine learning framework for enhancing the early detection of sepsis, leveraging the predictive potential of routine blood test metrics and Procalcitonin (PCT) levels. The study employs advanced data analytics to examine the interrelations between blood count behaviors and PCT, identifying subtle patterns that signal the onset of sepsis. Through rigorous algorithm testing and hyperparameter tuning, the research aims to develop a cost-effective and reliable diagnostic tool that significantly reduces the time to diagnosis. This approach promises to improve patient outcomes by enabling earlier and more precise interventions in sepsis treatment.

Keywords: Machine Learning, Sepsis Detection, Predictive Analytics, Blood Biomarkers, Hyperparameter Tuning, Diagnostic Accuracy, Blood Test Analysis, Healthcare Technology.

University Information:

University: University of Passau
Phone: +49 851 509 0
Address: Universität Passau, Innstraße 41, 94032 Passau, Germany