Mortality risk prediction for ICU septic shock patients using probabilistic fuzzy systems comparison of Sepsis-2 and Sepsis-3

Mustafayev, S.

Award date:
2017

Department of Mathematics and Computer Science
Department of Industrial Engineering & Innovation Sciences

Mortality Risk Prediction for ICU Septic Shock Patients Using Probabilistic Fuzzy Systems: Comparison of Sepsis-2 and Sepsis-3

Master's Thesis

Shamil Mustafayev
0980187

in partial fulfillment of the requirements for the degree of
Master of Science in Business Information Systems

Supervisors:
prof. dr. ir. Uzay Kaymak
dr. Anna Wilbik
Eindhoven University of Technology

prof. dr. Walther van Mook
dr. Dennis Bergmans
Department of Intensive Care
Maastricht University Medical Center+

Assessment committee member:
dr. Alexander Serebrenik

Eindhoven, August 2017

Abstract

Sepsis is a major problem and the main cause of death in intensive care units around the world. Despite advances in modern medicine and medical technologies, sepsis remains a leading cause of death. Overall mortality rates for sepsis neared 30% in 1979, but since 2000, these rates have decreased to 20%. The mortality rate for septic shock patients, however, is as high as 80%, which is much higher than the mortality of sepsis patients. In order to reduce the mortality rate of sepsis and septic shock patients, the severity of illness of such patients must be identified and treated immediately. Therefore, numerous mortality risk prediction tools were proposed to assist clinicians in the decision-making process. These prediction tools help clinicians in determining the severity of illness and, consequently, in aligning the type of treatment an ill patient receives. Probabilistic Fuzzy Systems, one of the mortality risk prediction tools, demonstrated satisfactory results in predicting the severity of illness. However, the definitions of sepsis and septic shock were updated in 2016; thus, the applicability of Probabilistic Fuzzy Systems based on these updated definitions has not yet been verified. Furthermore, the current literature lacks a well-structured approach for predicting mortality risk following the adoption of new definitions in this subdomain. In this master's thesis, we focus on septic shock patients and address this research gap by providing a comprehensive approach toward the adaptation of mortality risk prediction tools to newly established definitions. Then, using both the adapted tool and traditional machine learning techniques, we build a model to predict the mortality risk of a patient. Finally, we compare the performance of both the old and updated definitions and present the impact of the updated definitions on predicting the mortality rate of septic shock patients.

Keywords: Sepsis, SIRS, SOFA, Septic Shock, Sepsis-2, Sepsis-3, Mortality, Probabilistic Fuzzy Systems.

Acknowledgments

This master's thesis is the result of my graduation project for the Business Information Systems master's degree at Eindhoven University of Technology (TU/e). This work concludes my two-year academic journey at one of the best European universities.

First of all, I would like to express my deepest appreciation to my supervisor at TU/e, prof. dr. ir. Uzay Kaymak, for his support, dedication, useful comments, and engagement throughout the project. His guidance, critical assessment, and constant encouragement have helped me from the very beginning to the final stages of the project.

Second, I would like to thank my second supervisor, dr. Anna Wilbik, for her recommendations about the research process and the most relevant approach to take. This project would not have been possible without her suggestions and direction.

Third, I would like to thank my supervisors from Maastricht University Medical Center+ (MUMC+), prof. dr. Walther van Mook and dr. Dennis Bergmans. Their dedication, efforts, and knowledge helped me gain an understanding of the healthcare domain, specifically the sepsis subdomain, which was a mystery for someone like me without advanced medical knowledge. Thank you for your time during every productive meeting that we had.

Fourth, I would like to thank Jerry Townsend, a data manager at MUMC+, and prof. dr. Ronny Schnabel, an anesthetist-intensivist at the ICU of MUMC+ hospital, for helping me in identifying and retrieving the most relevant data necessary for my project. Thank you for your time and patience while sharing your medical knowledge with me.

Fifth, I would like to thank dr. Alexander Serebrenik for accepting the invitation to join my final assessment committee. Thank you for your time and interest.

I would also like to thank my best friend, Nderim Rexhepi. We started at this university together, we have shared our best moments over the past two years, and we have always supported each other along the way. Thank you for your sense of humor, support, and motivation.

Next, I would like to express my deepest gratitude to my parents for being supportive, inspirational, and caring. I would not have been able to complete this project without my lovely sister Narmin's encouragement as well. Thank you for being such an important part of my life!

Finally, I would like to thank every person who has helped me develop as a person and strive for excellence. Thank you, Sakina, Ilaha, Emin, Turgut, Javid, Kanan, Rufat and Sharif!

Shamil Mustafayev, August 2017

Contents

List of Figures

List of Tables

1. Introduction

This master's thesis is a graduation project for the Business Information Systems master's degree at Eindhoven University of Technology (TU/e) and has been accomplished in collaboration with Maastricht University Medical Center (MUMC).

Section 1.1 presents the thesis context by introducing the motivation behind mortality risk predictions within the Intensive Care Unit. The problem statement is formulated in Section 1.2 by emphasizing the need for prediction models in global healthcare. The research question of this project is stated in Section 1.3. The main question and sub-questions required to meet the objectives of this master's thesis and to address the gaps identified in the literature review are presented. Lastly, the overall research method and the outline of the subsequent chapters are described in Section 1.4.

1.1 Research Context

Sepsis is a major problem and the main cause of death in intensive care units (ICU) [3]-[4]. It is more common than heart attacks and results in more deaths than any cancer worldwide [9]. Kissoon et al. [5] claim that sepsis causes 60–80% of lost lives per year. Despite advances in modern medicine and medical technologies, sepsis remains a leading cause of death. The results of various research studies conducted by Benjamin et al. [6], Kaukonen et al. [7], and Shen et al. [8] indicate that the number of sepsis cases has increased steadily for the past 20 years. Overall mortality rates for sepsis neared 30% in 1979, but since 2000, these rates have decreased to 20% [9]. The mortality rate for septic shock patients, however, is as high as 80% [12], which is much higher than the mortality of sepsis patients. These mortality rates are considered unacceptably high [10]. In addition to the associated mortality, sepsis also represents a large financial burden on global healthcare. The Agency for Healthcare Research and Quality reported that with costs of over $20 billion, sepsis was the most expensive medical condition treated in U.S. hospitals in 2011 [11]. Such high mortality rates and treatment costs were observed in the Netherlands as well. In the first half of 2009, the mortality rate due to severe sepsis and septic shock in ICU patients in the Netherlands was 34.8% [71]. Various "surviving sepsis" campaigns were launched to increase the awareness about sepsis among the medical profession, governments, and the general public. Alternatively, numerous mortality risk prediction tools were proposed to assist clinicians in the decision-making process related to treatment and prognosis [34].

These prediction tools help clinicians determine the severity of illness and, consequently, customize the type of treatment an ill patient receives. This information helps clinicians in

making the choice of whether to intensify or reduce the treatment, which has an impact on saving resources (time, cost, etc.). However, what adds substantial value to this tool is that this information has a direct impact on selecting a treatment that can save a patient's life, making it a life-saving prediction tool [13].

In an extensive research study conducted by [72], accurate overall mortality prediction was observed in the mortality scoring systems, such as APACHE II, SAPS II, MPM II_0, and MPM II_{24}. However, since the definition and assessment of sepsis were both changed by Sepsis-3 in 2016, these predictions need to be remodeled. By making use of Probabilistic Fuzzy Systems (PFS), a type of classifier that has demonstrated satisfactory results in predicting the severity of illness [14], we can develop a model to forecast the mortality risk of septic shock patients. The goal of this project was to identify the possibility of model migration due to change of definitions and the consequences of that change on the modeling activities. The PFS-based model was used to generate the mortality risk indices of septic shock patients after 72 hours of admission based on the updated sepsis definitions.

1.2 Problem Statement

Following the above observations, the problem addressed in this thesis can be summarized as follows:

Problem Statement: There is no precise approach in model migration due to change of definitions in predicting the mortality risk of septic shock patients.

1.3 Research Questions

The objective of this thesis was achieved by answering the following main question and sub-questions:
How to migrate the model due to change of definitions and what are the consequences of that change on the modeling activities?

A. How should a predictive model be trained?

A final machine learning model is a model used to make predictions on new data. In our case, it is the most essential tool necessary for mortality risk prediction. This model is responsible for correctly classifying the outcome of a patient. Specific technique has to be identified and employed in order to train a prediction model and to obtain the best outcome. The final predictive model identifies the applicability of PFS for classification.

B. What is the impact of updated definitions on the performance of the PFS predictive model?

One of the main factors affecting the performance of the PFS predictive model is data. Specific data has to be gathered and processed that can be further used by the model. The final data is gathered and preprocessed using the updated definitions, thus having a vital effect affecting on the performance of the model. The impact of updated definitions may then be identified by comparing with the outcome of old definitions. By comparing the outcomes affected by these definitions, we may suggest clinicians to use the definition that had a better impact on the performance.

C. What are the requirements for predicting mortality risk if such changes occur in the future?

Due to advances in technology, the changes in a healthcare domain are inevitable. Therefore, it is utmost priority to know what steps need to be taken in order to adapt the new changes, and consequently analyze the worthiness of these definitions by building a model upon updated definitions.

D. What are the differences between the old model and the new model?

Changes in the definitions result in differences between the models. Knowing these dissimilarities, such as differences in variables and patient cohort, may help clinicians in understanding the essence of updated definitions.

1.3.1 Scope

The research focuses mainly on migration of the model due to adjustment of definitions. The mortality prediction of septic shock patients was performed using Probabilistic Fuzzy Systems in a Dutch setting based on old and updated sepsis definitions and was conducted within the Eindhoven University of Technology (TU/e). External data were obtained from Maastricht University Medical Center+ (MUMC+) as a part of the collaboration between TU/e and MUMC+. The hospital supervisors, as well as external supervisors, are dr. Dennis Bergmans and prof. dr. Walther van Mook, internists-intensivists.

1.4 Research Method and Outline

The following steps were taken in order to address the research questions:

- The general description of sepsis and its conditions have been described first. A detailed explanation can be found in **Chapter 2**. This chapter also contains a discussion about the existing decision support systems in the ICU and the use of Probabilistic Fuzzy Systems (FPS) for mortality prediction.

- **Chapter 3** presents the methodology followed in this thesis and complete implementation of Probabilistic Fuzzy Systems based on updated definitions to predict the mortality of septic shock patients.

- The results are presented in **Chapter 4,** where the differences between the models are also discussed.

- The conclusion of this thesis is presented in **Chapter 5** that also contains limitations of the proposed method and discussions about future work.

- A bibliography can be found at the end of the document.

2. Mortality Risk Prediction for Sepsis

A literature review is necessary to identify what has and has not already been investigated. Furthermore, it is important to identify the key concepts, data sources, and methodologies that are relevant to our research and have been previously investigated by other researchers. Hence, in order to address the problem statement, a thorough literature review was conducted. Firstly, the definition and prevention of sepsis are identified and presented, referring to various research papers. Secondly, sepsis-related decision support systems in the Intensive Care Unit are defined. Finally, the Probabilistic Fuzzy Systems used for predicting the mortality risk of sepsis and septic shock patients and the subsequent results are described.

2.1 Sepsis

The subsection starts with the introduction of the sepsis definition, mostly focusing on general information about sepsis. Lastly, the conditions of sepsis and its assessment criteria are presented in the level of detail that is necessary for mortality prediction.

2.1.1 Definition

The immune system of every person is strong enough to fight an infection in order to make the body healthy again. Sometimes, unfortunately, the body overreacts and causes tissue damage, organ failure, and death. This is known as sepsis and is defined as a life-threating organ dysfunction caused by a dysregulated host response to infection [1].

Until 2016, a diagnosis of sepsis required having two or more Systemic Inflammatory Response Syndrome (SIRS) criteria in combination with an infectious cause of SIRS [1]. Bone et al. [2] suggested that sepsis is a subcategory of SIRS, which is a serious condition related to systematic inflammation, organ dysfunction, and organ failure. However, in February of 2016, the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) was held to evaluate and update the definitions and criteria for sepsis and septic shock. The need for an update was necessary due to advances regarding unraveling the pathobiology (changes in organ function, morphology, cell biology, biochemistry, immunology, and circulation), management, and epidemiology of sepsis [1]. Proposed by Shankar-Hari et al. [1], SIRS does not reflect the actual existence of sepsis in ICU patients, since it is known that changes in white blood cell count, temperature, and heart rate reflect inflammation and that sepsis is now identified as comprising early activation of both pro- and anti-inflammatory responses [15]. Thus, SIRS was completely excluded from the definition of sepsis [1], resulting in redefinition for the states of sepsis. As known previously, Bone et al. [2] defined three altered states of sepsis: severe sepsis,

sepsis-induced hypotension, and septic shock. However, severe sepsis was claimed as being redundant by Shankar-Hari et al. [1] and, thus, only two states of sepsis were retained:

1. *Sepsis*: life-threatening organ dysfunction caused by a dysregulated host response to infection,

2. *Septic Shock*: subset of sepsis in which underlying circulatory and cellular/metabolic abnormalities are profound enough to substantially increase mortality.

The Third International Consensus Definitions for Sepsis and Septic Shock not only updated the definition of sepsis but also assessed new clinical criteria for sepsis and septic shock. The major change to criteria was the introduction of a Sequential [Sepsis-related] Organ Failure Assessment (SOFA) score for the identification of the condition. Complete definitions are in Table 1:

Organ dysfunction can be identified as an acute change in total SOFA score ≥ 2 points consequent to the infection. Note: The baseline SOFA score can be assumed to be zero in patients not known to have preexisting organ dysfunction.	**Respiratory system**		**Nervous system**	
	PaO2/FiO2 (mmHg)	**SOFA score**	**GCS**	**SOFA score**
	<400	1	13-14	1
	<300	2	10-12	2
	<200 and mechanically ventilated	3	6-9	3
	<100 and mechanically ventilated	4	<6	4
	Cardiovascular system		**Liver**	
	MAP or administration of vasopressors required	**SOFA score**	**Bilirubin (mg/dl) [µmol/L]**	**SOFA score**
	MAP < 70 mm/Hg	1	1.2-1.9 [>20-32]	1
	dop ≤ 5 or dob (any dose)	2	2.0-5.9 [33-101]	2
	dop > 5 or epi ≤ 0.1 OR nor ≤ 0.1	3	6.0 – 11.9 [102-204]	3
	dop > 15 or epi > 0.1 OR nor > 0.1	4	>12.0 [>204]	4
	Kidneys		**Coagulation**	
	Creatinine (mg/dl) [µmol/L] (or urine output)	**SOFA score**	**Platelets x 10^3/µl**	**SOFA score**
	1.2-1.9 [110-170]	1	< 150	1
	2.0-3.4 [171-299]	2	< 100	2
	3.5 – 4.9 [300-440] (or <500 ml/d)	3	< 50	3
	>5.0 [>440] (or <200 ml/d)	4	< 20	4

Table 1: SOFA scoring system

MAP – Mean Arterial Pressure

GCM – Glasgow Coma Scale
Dop – Dopamine
Dob – Dobutamine
Epi – Epinephrine
Nor - Norepinephrine

The Sequential Organ Failure Assessment (SOFA) score numerically quantifies the number and severity of failed organs. By the new definitions, an increase in the SOFA score of 2 points or more represents an organ dysfunction. A sepsis patient with such scores has an in-hospital mortality greater than 10% [58]. The complete description of the SOFA score is presented in 2.3.1.1.

In order to quickly identify patients with a suspected infection that are at greater risk for a poor outcome, new assessment criteria, known as Quick SOFA (qSOFA), was introduced. Poor outcomes are expected if at least two of the following qSOFA conditions are present:

1. Respiratory rate ≥ 22/min;
2. Altered mentation;
3. Systolic blood pressure ≤ 100 mmHg.

A patient is considered to have septic shock if sepsis with persisting hypotension requiring vasopressors to maintain MAP ≥ 65mmHg and a serum lactate level > 2 mmol/L (18 mg/dL) is observed despite adequate volume resuscitation [1].

In modern ICU settings, multiple variables of patients are recorded and available for measurement. By making use of this information, the values that are crucial for the identification of sepsis and septic shock patients can be extracted. Furthermore, we combine that information with demographics recorded upon admission to generate additional insights regarding the distribution of sepsis and septic shock mortality rates among age, year, length of stay, gender, etc.

2.2 Decision Support Systems in ICU

This subsection starts with a general description of decision support systems in the ICU, known as scoring models, to predict outcomes, along with exploring the disease severity of patients with sepsis. Next, widely used scoring models and the development of their versions are presented. Lastly, the prediction of mortality rates using Probabilistic Fuzzy Systems is reviewed.

2.2.1 Scoring Models

The first scoring model was developed more than 35 years ago and has since been the most common approach in assessing the severity of illness among critically ill patients and in predicting their mortality rate. These assessments improve clinical decision-making, which results in the improvement of patient outcomes. However, constant changes in intensive care practices, patient demographics, and disease pervasiveness [18] affected the continuous update of scoring models to provide more accurate results. Generic scores, such as Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), Mortality Probability Model (MPM), and Sequential Organ Failure Assessment (SOFA), are currently the most commonly used prediction models in the ICU. The selection of variables and their weights for these models used to be performed by a panel of experts. However, since 1985, the selection process has heavily relied on multiple logistic regressions. Common variables, such as age, surgical status, chronic health status, physiology, and acute diagnosis, are the main variables of such models. Upon admission to the ICU, risk-scoring models such as APACHE, SAPS, and MPM use the physiological values of the patient (within 24 hours) to determine the morbidity and mortality rates.

2.2.1.1 Methodology of Scoring Models

The abovementioned models have similar functionality. Initially, a nurse or a doctor obtains various physiological measurements, such as the systolic blood pressure, respiratory rate, and heart rate of a patient. These values are then converted to relevant points. Following the medical calculations by using each point, the total score is identified. As a result, this total score is used for determining the illness's severity and for predicting a patient's mortality rate. Although the methodology of scoring models is comparable, the variables being used for each model may vary. For instance, the SAPS II score uses 12 physiological variables and 3 disease-related variables, while APACHE II also uses 12 physiological variables but only 2 disease-related variables are being used for prediction. Moreover, unlike the SAPS II score, the APACHE II score can be converted to a mortality rate only if the reason for ICU admission is known. The complete description and calculation methods of each scoring model are described below.

SOFA

The Sepsis-related Organ Failure Assessment score, also known as SOFA, was developed in 1994 to assess the degree of organ dysfunction/failure of an individual or a group of patients. Unlike other scoring models, SOFA is not intended for predicting the mortality rate of a critically ill patient. The SOFA score is based on 6 different scores: respiratory, cardiovascular, hepatic, coagulation, renal, and neurological systems. A score from 0 (normal) to 4 (high degree of dysfunction) is assigned to each system depending on the measurement. The SOFA score range

is between 0 and 24. After assessing each physiological system separately, the total score is determined and the risk of morbidity is identified. The complete list of all variables, each representing an organ system, and relevant SOFA scores are described in Table 1.

Although SOFA is not intended for predicting the outcome, Vincent et al. [37] made a rough estimation of correlation between the total score and mortality risk. These estimations can be found in Table 2.

Maximum SOFA score	Mortality rate
0-6	< 10%
7-9	15-20%
10-12	40-50%
13-14	50-60%
15	>80%
15-24	>90%

Table 2: Correlation of maximum SOFA score with mortality rate

In 2008, a systematic review of studies evaluating the performance of Sepsis-related Organ Failure Assessment (SOFA)-based models was initiated by Minne et al. [73]. Eighteen articles related to predicting mortality in patients in the ICU using SOFA were selected for this study. The results of this review revealed that models based on SOFA scores at admission had better performance than SAPS II models. However, the authors suggest using the sequential SOFA derivatives with APACHE II/III and SAPS II models since this approach has demonstrated improved performance. In another study, this one conducted by Safari et al. [74] the accuracy of the SOFA score in predicting the 30-day outcome of critically ill patients was assessed. This study included 140 patients older than 18 years old. The outcome of the research concluded the fair accuracy of the SOFA scoring system.

APACHE I-IV

The first scoring system was developed in 1981. As described by Knaus et al. [20], the system, known as APACHE, is used to assess the degree of illness and to determine the chronic health status of patients by using physiological values. However, in 1985, a revised model, called APACHE II, was introduced, the major change of which was the reduction of physiological variables from 34 to 12. APACHE II is currently the most widely used scoring model. The maximum score of 71 can be obtained by assessing 12 routine physiological measurements ($AaDO_2$ or PaO_2, temperature, mean arterial pressure, pH arterial, heart rate, respiratory rate, sodium (serum), potassium (serum), creatinine, hematocrit, white blood cell count, and

Glasgow Come Scale). The worst 12 values of each measurement during the first 24 hours after the admission are utilized together with the age and chronic health status of the patient. The higher the score, the more severe the illness is. Although the previous health status of an ill person is not required for determining the severity of disease, it is one of the main criteria used in predicting the mortality risk. The morbidity and mortality rate of a patient may be calculated by using the following APACHE II equation:

$$x=-3.517+(0.146*points)+0.603(if\ emergency\ surgery)+(Admission\ Indication\ Weight)$$

$$ln(\frac{R}{1-R})=x \rightarrow \frac{R}{1-R}=e^{x} \quad (1)$$

$$R\ (percent\ mortality) = \frac{e^{x}}{1+e^{x}} * 100$$

APACHE III was introduced in 1991 [21] but was validated and updated in 1998 [22]. Similar to its predecessors, this model also uses the data of a patient collected during the first 24 hours of ICU admission. As described by Knaus et al. [20], the model also requires sex, race, preexisting comorbidities, and location prior to ICU admission. Unlike APACHE and APACHE II, this version uses logistic regression for the selection of variables and their weights. The range of an APACHE III score is between 0 and 299 points. The severity of illness is determined in the same way as previous APACHE models. A higher score means a higher acuteness and, respectively, a greater mortality risk. The last version of APACHE models, named APACHE IV, was the remodel of APACHE III and was developed in 2006. Although it uses the same physiological variables and weights as APACHE III, different predictor variables and refined statistical methods were incorporated in the final version of this scoring model [23]. In terms of performance, because the first developed APACHE was not intended for mortality prediction, numerous comparisons were conducted between the last three models of APACHE (II, III, IV). While comparing the performance of APACHE II and APACHE III in predicting the mortality of patients with sepsis, Sadaka et al. [26] concluded that both models perform equally well and both scores have very good discriminative powers. Further, Ayazoglu [24] compared the performance of APACHE II and APACHE IV scoring systems. His research indicated that these scores are equally good in predicting mortality rate in ICU stroke patients, but APACHE IV had a slight advantage over APACHE II in its reliability of prediction. In general, Abdelbaset et al. [25] suggest that each scoring system has a different aim and measurements. Hence, they should be perceived as complementary and not mutually exclusive.

MPM

The Mortality Probability Model, abbreviated as MPM, consisted of two different models: an admission model with seven admission variables and a 24-hour model with seven 24-hour variables [27]. Similar to previously mentioned scoring models, MPM required a revision as

well; hence, MPM II was developed in 1993. The major difference between these models was the introduction of a logistic regression technique [28]. Like its predecessor, MPM II also has two scores, namely, MPM_0 and $MPM_{24.}$ Unlike MPM models, MPM_0 consists of 15 admission variables. MPM_{24}, conversely, consists of 5 admission and 8 additional variables. Lemeshow et al. [27] claim that MPM_0 is the only available model that may be used upon ICU admission. Although most well-known scoring models weight the variables, MPM II variables are labeled as either present or absent and are given a score of 1 or 0 respectively. After obtaining the required variables, a probability of hospital mortality is calculated by using a logistic regression equation. Additionally, to gain insight into resource utilization, Rapoport et al. [29] developed a Weighted Hospital Days scale (WHD-94), where the weights were assigned to days in the ICU and to days after the discharge from the ICU. The WHD-94 equation is capable of predicting resource utilization by calculating an ICU's mean. To assess clinical performance, a revised MPM_0, MPM_0-II, was also developed by Rapoport et al. [29]. However, these models do not calibrate on contemporary data. Thus, both MPM_0-III and WHD-94 models were updated in 2007 [30], [31] to estimate mortality probability and resource utilization respectively. MPM_0, MPM_{24}, $MPM_{48,}$ and MPM_{72} models are part of Mortality Probability Models and are used for respective time periods from ICU admission. In terms of performance, except for the MPM_0 model, other models have shown good calibration and discriminative powers [30], [31], [32].

SAPS

The Simplified Acute Physiology Score, simply SAPS, was developed in 1984 and was aimed at predicting the risk of death in ICU patients [33]. SAPS was calculated using 13 weighted physiological variables, and similar to APACHE, it used the worst 12 values obtained during the first 24 hours after admission to the ICU. In 1993, however, Le Gall et al.[13] developed a newer version of this scoring model, known as SAPS II, where 12 physiological variables, 3 underlying disease variables, age, and type of admission were used for logistic regression analysis to build this model. The score was validated using 12,997 patients from 137 ICUs in 12 countries [33]. Although SAPS II was found to be superior to SAPS, a study conducted by Metnitz et al. [36] found it to be poorly calibrated. This model's score range is between 0 and 163 points. Like APACHE, SAPS II weights were estimated using multiple logistic regression. Total score and hospital mortality rate may be calculated using (2).

$$logit=-7.7631+(0.0737*score)+0.9971*ln(score+1) \quad (2)$$

$$Mortality = \frac{e^{logit}}{1+e^{logit}}$$

In 2005, a completely refined model of SAPS, the SAPS 3, was created. Unlike other scoring models, SAPS 3 uses 20 variables divided into three sub-scores: Patient characteristics prior to admission, the reason for admission, and the worst values measured within one hour before or

after ICU admission. The score ranges from 0 to 217. SAPS 3 equations are customized based on the locality of patients, which is the main advantage over other risk-scoring models. Using this model, the mortality rate of 6 regions—Australasia; Central and South America; Central and Western Europe; Eastern Europe and Northern Europe; Southern Europe and the Mediterranean; and North America—may be calculated. Similar to its predecessor—SAPS II—weights for SAPS 3 were assessed using multiple logistic regression [34]. In research conducted by Moreno et al. [75], SAPS 3 has demonstrated good discrimination, calibration, and goodness of fit.

qSOFA

Quick SOFA, simply known as qSOFA, is a new scoring system introduced at the 2016 Third International Consensus Definitions for Sepsis and Septic Shock. It is aimed toward the identification of patients with suspected infection who are more likely to have a poor outcome outside the ICU. Moreover, qSOFA simplifies the SOFA score by only having 3 clinical criteria. The qSOFA score ranges from 0 to 3. The assessment of qSOFA and its corresponding score is given in Table 3.

Assessment	qSOFA score
Low blood pressure (SBP ≤ 100mmHg)	1
High respiratory rate (≥ 22 breaths/min)	2
Altered mentation (GCS < 15)	3

Table 3: qSOFA assessment

For instance, a patient with low blood pressure (e.g. 60mmHg) and altered mentation (GCS=12) would score 4 qSOFA points. A qSOFA score of ≥ 2 proposes a 3- to 14-fold increase in the rate of in-hospital mortality [58]; therefore, the presence of organ dysfunction in such patients should be assessed by calculating a SOFA score as well. However, Vincent et al. [37] claim that a patient may have a qSOFA ≥ 2 without infection, or sepsis may be present without a qSOFA score ≥ 2 in numerous acute conditions. Thus, quick SOFA has raised many concerns among scientists and is considered an imperfect marker of sepsis.

2.3 Probabilistic Fuzzy Systems for Mortality Prediction

The term *fuzzy* and fuzzy-related concepts were proposed by Zadeh [43]. Fuzzy logic is a computing approach where the “degrees of truth” may be any real number between 0 and 1, unlike Boolean logic, where the variables take only “true or false” (1 or 0) values. Because fuzziness and probability are distinct phenomena, they should be accepted as complementary [38]. The first use of fuzzy logic in clinical decision support systems was proposed by Warren et

al. [76], but uncertainty concerning the occurrence of an event in the future required a different approach. Thus, in 2011, Meghdadi et al. [39] recommended Probabilistic Fuzzy Systems (PFS) to model vagueness in linguistic terms (fuzziness) and probabilistic uncertainty simultaneously.

The first use of PFS for classification to predict the mortality of ICU septic shock patients was initiated by Fialho et al. [14]. A total of 131 patients admitted to the ICU with abdominal septic shock were selected for the experiment. The following physiological values, including systolic blood pressure, platelets, spO_2, white blood cell count, arterial pO_2, and PTT, were carefully chosen based on the previous process of feature selection [40]. Data preprocessing was initiated in order to improve the quality of the data. For instance, the difference in sampling periods urged the necessity of choosing a template variable. Heart rate was selected as a template variable since it was the most frequently measured variable, with one sample per hour. In terms of model setup, two models, namely, Fuzzy Systems and Probabilistic Fuzzy Systems, were trained and tested on equally-sized subsets. Then the fuzzy c-means clustering approach, with two parameters (number of clusters and the degree of fuzziness of the clustering), was used for PFS modeling. As the final step, the model assessment was performed and the discrimination, sensitivity, specificity, and accuracy of the model were obtained. To measure the difference in performance, first-order Takagi-Sugeno Fuzzy Models (FM) were applied to the same data set [41]. As a result, both FPS and FM had good results in all assessment criteria. The complete results of the research are presented in Table 4.

	PFS	FM
AUC	0.80 ± 0.02	0.81 ± 0.03
Specificity	0.79 ± 0.03	0.79 ± 0.02
Sensitivity	0.81 ± 0.04	0.82 ± 0.03
Accuracy	0.78 ± 0.03	0.80 ± 0.02

Table 4: Results of Probabilistic Fuzzy Systems and Fuzzy Models [14]

The latest research of Probabilistic Fuzzy Systems for mortality prediction was also carried out by Fialho et al. [64]. In total, 4 models, specifically the Probabilistic Fuzzy Systems with maximum likelihood estimation (PFS-ML) model, Probabilistic Fuzzy Systems with conditional probability (PFS-CP) model, Fuzzy Model (FM), and Logistic Regression (LR) model, were assessed. The primary goal was to calculate the patient condition (alive or deceased) in a 72-hour window; hence, 27 variables representing 81 septic shock patients were selected. Briefly, Fuzzy Modeling, Probabilistic Fuzzy Systems, Sequential Method, maximum likelihood estimation, feature selection, model assessment, and simplification of fuzzy rules were implemented consecutively to obtain and compare the final results. Out of 81 patients, 36 were deceased, which resulted in a 42.3% mortality rate. Regarding classification performances,

except PFS-CP, all three types of models have demonstrated satisfactory results. Concerning the selected variables, each modeling technique has selected a different number of features; the most common features are diastolic blood pressure, platelets, spO_2, and FiO_2. Finally, the results in Table 5 were obtained after the assessment of all models based on most predictive variables through 10-fold cross validation.

	PFS-ML	PFS-CP	FM	LR
AUC	0.81 ± 0.02	0.68 ± 0.03	0.83 ± 0.03	0.83 ± 0.02
Specificity	0.81 ± 0.03	0.67 ± 0.02	0.82 ± 0.03	0.83 ± 0.03
Sensitivity	0.78 ± 0.03	0.65 ± 0.03	0.82 ± 0.02	0.85 ± 0.02
Accuracy	0.83 ± 0.04	0.70 ± 0.02	0.84 ± 0.03	0.81 ± 0.03

Table 5: Results of PFS-ML, PFS-CP, FM, and LR models [64].

From the results of Fialho's works, we can conclude that Probabilistic Fuzzy Systems demonstrate satisfactory results in Mortality Prediction. Therefore, Probabilistic Fuzzy Systems were employed to reach the objectives of this master thesis.

3. Mortality Prediction Modeling with Fuzzy Systems

This chapter starts with the introduction of Fuzzy Modeling and the type of fuzzy system, namely, the Probabilistic Fuzzy System (PFS). Later, theoretical analysis of selected feature selection, model assessment, and simplification of fuzzy rules are presented. The general description of chosen data sets, applied data preprocessing techniques, and details of the experimental setup are presented in this chapter as well.

3.1 Modeling

The basics of the selected modeling technique used to meet the objectives of this thesis are presented in the following subsections. In mortality prediction, the output is considered as a binary problem with two classes. If *c* is denoted as an output and *C* as its class, $c \in \{C_1, C_2\}$, where each C corresponds to the classes "patient will survive" (0) and "patient will die" (1) respectively.

3.1.1 Fuzzy Modeling

The concept of a set and set theory are powerful concepts in mathematics. As described by the principal notion underlying the set theory, an element can exclusively belong or not belong to a set. However, it is almost impossible to represent many elements of human speech, such as *short* distance, *weak* person, and *talented* student. Thus, Zadeh [43] introduced the fuzzy set theory to deal with uncertainty as regular sets with sharp boundaries. This approach is very powerful since "it provides not only a quantitative model, but also a linguistic interpretation in the form of rules and logical connectives" [14]. Contrary to clinical set theory, elements in fuzzy sets are not seen as being either part of it or not but rather being similar to other elements described by a fuzzy set. The numerical (membership) value assigned to elements varies from 1 (equal) to 0 (dissimilar).

In order to describe an element by assigning a membership value, sets of if–then rules with a fuzzy inference mechanism is used. The collection of simple if–then rules is part of fuzzy logic systems, which take a number of fuzzy sets as inputs (premises) and create fuzzy sets as outputs (conclusions). Such if–then rules can easily be interpreted by a human reader. This graduation project uses first-order Takagi-Sugeno (TS) Fuzzy Interference Systems [60] to classify each instance of a data set to one of the predefined classes. Takagi-Sugeno rules are defined as follows:

$$\text{Rule } R_j^c: \text{If } x_1 \text{ is } A_{j1} \text{ and ... and } x_M \text{ is } A_{jM}$$
$$\text{Then } d_c(x) = f_j^c(\mathbf{x}),\ j = 1, 2, ..., N, \tag{3}$$

where f_j is the consequent function (output) for rule R_j, M is the total number of inputs (features) and the index c indicates that the rule is associated with the output class c [44]. Considering the fact that predicting mortality risk is a classification problem, the continuous output $c' \in \{0,1\}$ should be converted into the binary output $c \in \{0,1\}$. Thus, the threshold t is required to carry this operation. To satisfy the abovementioned rules, if c'<t, then c = 0, and if $c' \geq t$, then c = 1.

3.1.1 Probabilistic Fuzzy System

A Probabilistic Fuzzy System consists of a set of rules whose antecedents are fuzzy conditions and whose consequents are probability distributions. This fuzzy system was developed based on the "probability of a fuzzy event" concept suggested by Zadeh [43].

Consider the task of determining the crisp class $c \in \{C_1, ..., C_c\}$ to which a data point $\mathbf{x} = (x_1, ..., x_d) \in X$ belongs. Since it is a binary problem, the probabilistic fuzzy classifier has the following form [44]:

$$\text{Rule } R_j: \text{If } \mathbf{x} \text{ is } A_j \text{ then } c = C_1 \text{ with probability } p_{j,1} \text{ and}$$
$$c = C_2 \text{ with probability } p_{j,2} \tag{4}$$

where $j = 1, ..., J$ corresponds to the rule number, A_j is the antecedent fuzzy set of the j^{th} rule, $p_{j,1}$ and $p_{j,2}$ are the probability parameters of patients surviving (C_1) and patients dying (C_2) classes, respectively, for rule j.

The antecedent fuzzy sets A_j in (4) are defined in the d-dimensional input space X. It is important for each fuzzy set to correspond to a probabilistic fuzzy rule. Furthermore, the probability parameters $p_{j,k}$ in (4) satisfy

$$p_{j,k} \geq 0 \quad \text{for} \quad j = 1, ..., J \text{ and } k = 1, ..., c \tag{5}$$

and

$$\sum_{c=1}^{C} p_{j,k} = 1 \qquad \text{for } j = 1, ..., J. \tag{6}$$

Let $\mu_{Aj}(\mathbf{x})$ denote the membership function of a fuzzy set A_j. A probabilistic fuzzy classifier with rules given by (4) was used to estimate $\Pr(c \mid \mathrm{x})$, the conditional probability distribution of c given x. As described in Takagi-Sugeno fuzzy reasoning, the normalized membership functions μ_{Aj} define the probabilistic fuzzy rules activations. Therefore, the estimate $\hat{p}(C_k \mid \mathrm{x})$ of a conditional probability $\Pr(C_k \mid \mathrm{x})$ is obtained as follows [45]:

$$\hat{p}(C_k \mid \mathbf{x}) = \sum_{j=1}^{J} \mu_{Aj}(\mathbf{x})\, p_{j,\,k.} \tag{7}$$

where $\beta_{Aj}(\mathbf{x})$ indicates the normalized rule activation:

$$\beta_{Aj}(\mathbf{x}) = \frac{\mu_{Aj}(\mathbf{x})}{\sum_{j=1}^{J} \mu_{Aj}(\mathbf{x})}. \tag{8}$$

Given $\mathbf{x}$, $\beta_{Aj}(\mathbf{x})$ denotes the degrees of fulfillment of each of the probabilistic fuzzy rules. Obtained as a weighted average of the rules that fire, an estimate of the probability is provided by (7). Certainly, as with traditional fuzzy systems, the accuracy of the estimate is correlated with the number of rules. A higher number of rules provide better accuracy of the estimate.

Furthermore, the estimated conditional probability distribution $\hat{Pr}(c \mid x)$ can be used in order to classify a data point x. The probability of misclassification may be minimized by the following classification rule:

$$\hat{C} = \underset{c \in \{C_1, C_2\}}{\arg\max}\, \hat{Pr}(c \mid \mathbf{x}) \tag{9}$$

Thus, in a probabilistic fuzzy classifier, two conditions must be identified: Firstly, the parameters and type of the antecedent membership functions; secondly, the probability parameters of the rule consequents.

The parameters of the probabilistic fuzzy classifiers can be determined in several ways. Kaymak et al. [43] proposed a sequential method for parameter estimation in a probabilistic fuzzy classifier. The approach is as follows. The parameters are divided into two disjoint sets, and each set of parameters is estimated. Usually, the first set of parameters is used for determining the parameters of the antecedent membership functions by using unsupervised learning methods or expert knowledge. Furthermore, assuming that the parameters in the first set are constant, the probability parameters for the rule consequents are estimated using a formula based on the probability measures of fuzzy events.

Waltman et al. [41] suggested that the sequential parameter estimation method does not provide optimal estimates of the parameters in a probabilistic fuzzy classifier. The main drawback of this approach is described as the parameters not being estimated simultaneously, where the optimal parameters in one set of parameters may not be optimal for the entire classifier. Additionally, the performance of the classifier is negatively affected by the fact that unsupervised learning of the antecedent parameters does not take class labels into account. Yet another issue with this approach is that the Conditional Probability (CP) method does not maximize the probability of observing the data set available for parameter estimation. Therefore, Waltman et al. [41] proposed a Maximum Likelihood Estimation (MLE) method that is superior to sequential parameter estimation. A proposed MLE method was used to estimate both the antecedent parameters and the probability parameters for the probabilistic fuzzy classifier while maximizing the likelihood of the data set available for parameter estimation and estimating all parameters simultaneously to meet the identification objective.

However, Fialho et al. [14] demonstrated that the combination of Sequential Parameter Estimation and Maximum Likelihood Estimation methods provide satisfactory results. Hence, in this work, sequential parameter estimation is used for finding the initial values of the parameters, following the optimization of these values by applying the MLE method.

3.1.1.1 Sequential Method for Parameter Estimation

In order to estimate the rule and probability parameters, these parameters first have to be initialized and then estimated. The initial estimation of the rule parameters is done using Fuzzy C-means Clustering (FCM). Furthermore, the conditional probability method is used to initialize and estimate the probability parameters. Finally, the maximum likelihood estimation method is applied to optimize these values.

A sequential method for parameter estimation approach is taken from the paper of Waltman et al. [41]. In their work, it is assumed that each antecedent membership function $\mu_{Aj}(\mathbf{x})$ in a probabilistic fuzzy classifier is the product of d univariate Gaussian membership functions $\phi(x) = \exp(-(x-c)^2 / \sigma^2)$, one for each dimension of the input space X. Following the abovementioned assumption, the following result is obtained:

$$\mu_{Aj}(x) = \exp(-\sum_{l=1}^{d} \frac{(x_1 - c_{jl})^2}{\sigma_{jl}^2}). \tag{10}$$

For each membership function, $c_j = \{c_{j1,\ldots,} c_{jl}\}$ and $\sum_j$ = $\{\sigma_{j,1,} \ldots, \sigma_{j,l}\}$ vectors represent the parameters that need to be estimated. The vectors c_j and $\sum_j$ correspond, respectively, to the center and the width of the membership function in each dimension of input space *X*. After the

estimation of c_j and $\sum_j$ vectors, the probability parameters p_{jk}, while satisfying (5) and (6), also need to be estimated. In the sequential method for parameter estimation, the antecedent parameters are determined separately from the probability parameters p_{jk} [41].

For the estimation of antecedent parameters c_j and $\sum_j$, the data set containing n classification examples (x_i, y_i) (i = 1, .. ., n) must be available. Firstly, this data set is normalized using the following formula:

$$\tilde{x}_{il} = \frac{x_{il} - \mu_l}{\sigma_l} \tag{11}$$

where μ_l and σ_l denote, respectively, the mean and the standard deviation of the l^{th} feature over the entire data set. Furthermore, for the set number estimation of cluster centers, the fuzzy c-means algorithm is applied to the normalized observations $\tilde{x}_i$. The FCM uses the standard Euclidean distance measure, while obtained cluster centers serve as the centers of c_j of the Gaussian membership functions μ_{Aj}. After obtaining the cluster centers, the vectors $\sum_j$, containing the width of the membership functions must be estimated. In order to estimate these vectors, the nearest neighbor heuristic [46] approach is used. The following approach results in:

$$\sigma_{j,l} = \min_{j' \neq j} \| c_j - c_{j'} \| \quad \text{for } l = 1, \ldots, d \tag{12}$$

where $\| c_j - c_{j'} \|$ represents the Euclidean distance between c_j and $c_{j'}$. It is important to note that the width of a membership function is the same in each dimension.

As discussed earlier, the probability parameters p_{jk} need to be estimated. Kaymak et al. [44],[45],[47] suggest setting these parameters as equal to estimates of the conditional probabilities Pr(C_k|A_j). Proposed by [47], this approach results in:

$$p_{j,k} = \frac{\sum_{i=1}^{n} \tilde{\mu}_{A_j}(x_i) \chi_{C_k}(y_i)}{\sum_{i=1}^{n} \tilde{\mu}_{A_j}(x_i)}, \tag{13}$$

where $\chi_{C_k}(y)$=1 if y=C_k and $\chi_{C_k}(y)$=0 otherwise. Equation (13) is usually referred to as the *conditional probability estimation.* Waltman et al. [41] suggest that this estimation is biased and that maximum likelihood estimation offers better estimates of the parameters.

3.1.1.2 Maximum Likelihood for Parameter Estimation

After presenting the drawbacks of the sequential method and the advantages of maximum likelihood for parameter estimation in section 3.1.1, this section contains the mathematical description of the maximum likelihood approach.

The likelihood of a data set is given by

$$L = \prod_{i=1}^{n} p(y_i \mid x_i). \tag{14}$$

Maximization of the likelihood is equivalent to the minimization of the negative log likelihood [41]. Therefore, the following error function is minimized:

$$E = -\sum_{i=1}^{n} \ln \hat{p}(y_i \mid x_i). \tag{15}$$

In order to achieve the minimization of the error function in (15), a gradient descent optimization algorithm is used. The stochastic variant of gradient descent is applied, which means that the available classification examples are processed one by one, and updates are performed after each example [41]. In order for a gradient descent optimization algorithm work, we have to define the learning rate to an appropriate value. This value determines in what pace we are moving towards the optimal weights. If the learning rate is very high, the optimal solution may be skipped. If the learning rate is too low, then it is necessary to run many iterations to converge to the best values. Thus, using a decent rate is crucial.

Subsequently, the probability parameters p_{jk} must satisfy the conditions in (5) and (6), finding the parameters c_j, $\sum_j$ and $p_{j,k}$ that minimize the error function in (15) is a constrained optimization problem. Thus, this constrained optimization problem is converted into an unconstrained optimization problem by using the auxiliary variables $u_{j,k}$ ($j = 1, \ldots, a$ and k = 1, ...,c). The relation between these variables and the probability parameters $p_{j,k}$ is described by the *softmax* function, i.e.,

$$p_{j,k} = \frac{e^{u_{j,k}}}{\sum_{k'=1}^{c} e^{u_{j,k'}}} \tag{16}$$

Maximum likelihood estimates of the parameters c_j, $\sum_j$ and p_{jk} can then be obtained by the unconstrained minimization of (15) with respect to c_j, $\sum_j$ and u_{jk}.

Algorithm 1 is proposed by Waltman et al. [41] and summarizes the Maximum Likelihood method to determine the parameters of a probabilistic fuzzy system:

Algorithm 1: Maximum Likelihood method to determine parameters c_j, $\sum_j$, and $p_{j,k}$ of a Probabilistic Fuzzy System.

1) **Initialize** c_j, $\sum_j$, and $p_{j,k}$
   - Apply FCM clustering to find c_j, and $\sum_j$
   - Determine initial values of $p_{j,k}$ by, $p_{j,k} = \frac{\sum_{i=1}^{n} \tilde{\mu}_{A_j}(x_i)\chi_{C_k}(y_i)}{\sum_{i=1}^{n} \tilde{\mu}_{A_j}(x_i)}$

2) **Maximum likelihood**

   for *iter* = 0 to *no. of iterations* **do**

      for *train* = 1 to *no. of classification examples* **do**

         - Simultaneous re-estimation of the parameters v_j, $\sum_j$, and $p_{j,k}$ through minimization of error function (15)

      end for

      - Determine $E(iter) = -\sum_{i=1}^{n} \ln \hat{p}(y_i \mid x_i)$.

   end for

3.1.2 Feature Selection

Features are the main determinants in model construction. Usually features are numeric, but they can also be strings and graphs. However, it is expected that the greater number of features increases the accuracy of the model; for some classifiers, it negatively affects the reliability of statistical parameter estimations, and thus it results in a decrease in the classification accuracy [48]. This phenomenon is known as the Hughes Effect, which describes the correlation between the number of features and the accuracy of classification. Hughes [49] suggested that the feature set size may be proportionally increased to the number of training samples. Otherwise, the accuracy of classification will decrease once the feature set size is reached. In our work, the feature selection algorithm was employed to select the most relevant features for mortality risk prediction.

Feature selection methods can be used to identify and remove unnecessary, irrelevant, and redundant variables from a data set that do not improve the accuracy of a predictive model or may even negatively affect the accuracy of the model. In terms of preferred algorithms, feature selection algorithms can be divided into four different categories: filters, wrappers, hybrids, and embedded [50], [51]. The simplest type of feature selection algorithms is the filter method. This method selects the feature subsets without employing any mining algorithm simply by relying

on general characteristics of the training data to select features with independence of any predictor. Filters, such as chi squared test, information gain and correlation coefficient scores, provide a generic selection of variables and those variables are not tuned for a given learning machine. Somol et al. [65] concluded that the filters performed very poorly due to weak principal connection between the search and classification processes. Unlike filter method, the wrapper method employs one mining algorithm and uses the performance results for evaluation purposes. Many wrapper methods, such as best-first, branch-and-bound, genetic algorithms, forward selection, or backward elimination, are used to search for the features that improve the performance of the chosen algorithm. While comparing the performance of wrapper and filter methods, Somol et al. [65] concluded that the results of their research show an overall superiority of wrappers over filters. The hybrid method, however, takes advantage of the two previously explained methods. Firstly, a filter approach is applied to eliminate features according to the predefined criteria. Next, as a part of the wrapper method, a mining algorithm is used to select the relevant features. In contrast to previous feature selection algorithm categories, in the embedded method, the learning and feature selection parts are processed together. Decision trees are part of the embedded method. Overall, it is suggested to employ the hybrid methods with a large data set, whereas the wrapper methods have demonstrated better results with a small sample size [66].

After careful analysis of various feature selection methods, a wrapper method, namely sequential forward selection (SFS), was used in this project. Unlike other methods, SFS selects only the most relevant features and uses the performance results for evaluation purposes. Sequential forward selection, also known as bottom up selection, is an approach that starts with an empty feature set and adds one feature at a time. After the first feature is added to the subset, a new model is trained and the performance is tested. The next feature is then added to the same subset, and the same process is repeated until the optimal feature subset is obtained and adding a new feature does not improve the performance. Typically, the most occurred feature (variable) is added to the subset first, and the selection of consequent variables depends on the performance of the model. As the performance criteria, Area Under the Curve (AUC) is selected and used for the model evaluation. Although the sequential forward selection approach is considered simple and easy for graphical representation and the interpretation of results, the main disadvantage is known to be finding the local optima in a greedy and thus prone method [40].

For instance, if we applied sequential forward selection to MIMIC data set, which comprises over forty thousand patients who stayed in critical care units of Beth Israel Deaconess Medical Center between 2001 and 2012, we would start with an empty data set. Then, we would feed a single variable (e.g. glucose) to the model and compute the performance of this model upon the test set (e.g. an AUC of 0.63). Then other feature candidates are evaluated, one at a time,

and the feature that returns the best value of the performance criterion becomes the first selected feature. At the next step, other feature candidate, say, creatinine is added to the previous best model. This time the model is trained and tested on the combination these two features. Yet again, one feature at a time is added to the previous best model and the model is evaluated. The model with the highest performance is selected and now it has two features. This process is repeated until the value of performance criterion (e.g. AUC) stops increasing. In our thesis, the AUC is selected as the performance criterion and is described in the following section.

3.1.3 Model Assessment

Several model performance assessment methods exist to evaluate the performance of a classifier. Although the traditional confusion matrix or contingency table is used for the assessment of classification model performance, Fawcett et al. [77] suggest that the performance metrics derived from these matrices or tables are sensitive to data anomalies, such as class skew. Thus, David et al. [78] proposed using Receiver Operating Characteristics (ROC) curves that deliver the same information but in a much more intuitive and robust way. Subsequently, since we are looking into the classification problem with predicting survival outcomes, the area under the receiver-operating characteristic curve (AUC) can be used to measure how well a parameter can distinguish between two diagnostic groups (deceased/survived) [52]. AUC has a meaningful interpretation for disease classification in a healthcare domain. The main advantage of the area under the curve is the ability to determine the optimal cut off. Last but not least, this curve plays a key role in "evaluating diagnostic ability of tests to discriminate the true state of subjects" [67].

3.1.3.1 ROC Curves

ROC curves are two-dimensional graphs that visually represent the performance and performance trade-off of a classification model [79]. In order to construct ROC curves, the confusion matrix (or contingency table) should be created. Since the classification model classifies each instance into one of two classes, namely, a true (deceased) and a false (survived) class, this results in four possible classifications for each instance: a true positive (TP), a true negative (TN), a false positive (FP), or a false negative (FN). The format of a confusion matrix is depicted in Figure 1, where the columns and the rows represent the observed and the predicted classifications, re

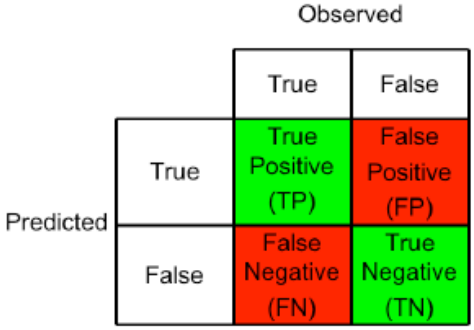

Figure 1: Format of a Confusion Matrix

The most common model performance metrics derived from the Confusion Matrix are accuracy (17), precision (18), and recall (19).

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (18)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (19)$$

To compare two different models, since it is more convenient to have a single metric rather than several, two new performance metrics are computed and later combined into one:

True positive rate ($\frac{TP}{TP+FN}$) reflects the proportion of correctly predicted positive cases over the total number of positive cases. False positive rate ($\frac{FP}{FP+TN}$), however, corresponds to the proportion of incorrectly predicted positive cases over the total number of negative cases. The true positive rate (TPR), known as sensitivity, and (1-false positive rate), known as specificity of the model, represent the cases where the patient was correctly classified as dying and surviving, respectively. After computing TPR and FPR, the ROC curve can be plotted. Each point on the ROC curve represents a sensitivity/specificity pair that illustrates the particular decision threshold. In our case, the sensitivity and specificity of the model corresponds to the cases where the patient was correctly classified as deceased and the cases where the patient was correctly classified as survived, respectively. Furthermore, the AUC, which is a single metric, can be used to assess the discriminatory power of the model [52].

3.1.4 Cross-validation

In 3.1.3, different types of metrics were described for assessing the model. Those metrics are used to evaluate model performance regarding our in-sample data. However, a specific model validation technique, namely, the cross-validation technique, is selected to assess how the results of a statistical analysis will generalize to an independent data set. Two types of cross-validation can be performed: leave-one-out and k-fold. Leave-one-out, however, is computationally demanding. Thus, the k-fold cross validation technique is used in this work.

In the k-fold cross-validation approach, each record is used k times for training and exactly once for testing. Briefly, suppose the data is being divided into two equal-sized subsets. First, one of the subsets is chosen for training and the other for testing. Then the roles of the subsets are swapped. This approach is called a two-fold cross-validation. The total error is acquired by summing up the errors of both runs. In this project, however, a 10-fold cross-validation was employed. Similar to the two-fold approach, in 10-fold cross-validation, the data is divided into

10 equal-sized portions. In each run, one of the portions is selected for testing and the rest are used for training. This process is repeated 10 times so that each portion is used for testing exactly once. Finally, the total error is computed by summing up the errors of all 10 runs, while AUC, sensitivity, and specificity results were obtained by averaging each result over the runs.

3.2 Experimental Setup

In order to test the hypothesis and reach the objectives of this project, an experiment is conducted. At the beginning, unaltered data sets are acquired. In this project, data obtained from MUMC+ was used. Then data preprocessing methods, such as the removal of outliers, as well as applying inclusion and exclusion criteria, were applied to the data sets. Following these preprocessing methods, the data was ready to be used for mortality prediction. Furthermore, the ready data set was split into 80% for feature selection and model building with the remaining 20% for model validation. Both data sets were undersampled to contain at least 48% observations with 72-hour mortality as an outcome [14]. Since each observation was labeled with 0s and 1s, the minimal percentage of 1s within the data set was set to 48%. This approach is necessary to prevent the misclassification due to a high ratio of surviving patients over deceased patients. The performance of the model is verified using 10-fold cross-validation. Figure 2 represents a complete diagram of this model setup.

As a part of model settings, the following parameters were used: The probability measures of PFS were calculated using MLE estimation. The number of rules used for the analysis was set to two since it showed satisfactory results. These rules correspond to the number of clusters obtained from the FCM-based clustering process. Furthermore, the Gaussian membership function was used to represent vague and linguistic terms. This function requires the widths to be determined in advance, thus $\sigma_{nj\geq} 0.1$ with, (j=1,…,J) and (n=1,…,N), was set. Such an approach was also necessary to avoid numerical instability. The number of learning iterations (Algorithm 1) and the learning rate, however, were set to 50 and 0.03, respectively. The number of iterations was fixed to a number that guaranteed that the error (15) would converge [41]. Thus, the use of convergence criteria was eliminated from the process [14].

Algorithm 2: Experimental setup [63]

Require: Original unaltered data set

1) Start data preprocessing

a. Include patients using inclusion criteria

b. Remove unusable features

c. Merge duplicate feature IDs

d. Clean data using pre-defined criteria

Return: Cleaned data set

2) Split data into 80% Feature Selection & Model Building and 20% Model Validation
3) Perform feature selection to obtain optimal feature subset
4) Perform 10-fold cross-validation to determine optimal model parameters
5) Perform model validation on holdout sample
6) Perform external validation on samples from other data sets
Return: Optimal model with internal and external performance measures

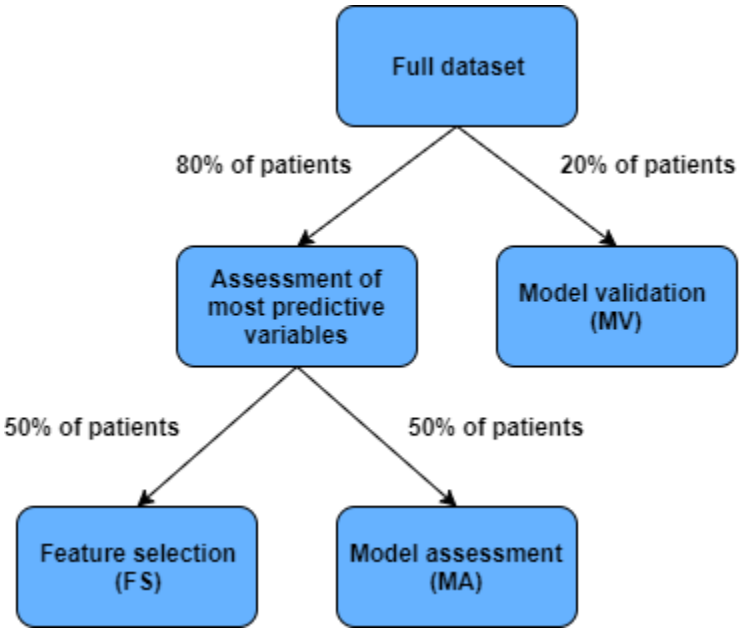

Figure 2: Diagram of the model setup [14]. Each data set (Sepsis-2 and Sepsis-3) follows the same model setup.

3.2.1 Data Preprocessing Technique

In order to improve the quality of the data, few data preprocessing techniques are usually employed. One of those techniques is related to missing data, since it greatly affects the quality of data. Missing data can significantly affect the performance of predictive risk modeling, which is an important technique for developing medical guidelines [55]. Imputing or deleting the missing values are the two most commonly used strategies for resolving the missing data issue. Although the first strategy may cause bias, the second strategy causes both bias and a loss of statistical power [55].

The built model requires the data to be presented in such a format that every row represents the patient ID, the date, the time, and all the features that are recorded for the same patient at a given moment. Even though the features are recorded automatically using medical machines, not all features are collected at the same time. For example, blood pressure may be measured

hourly, while blood gasses may be measured every 2 hours. A gridding template with a 1-hour sampling frequency will therefore show many instances of missing data for blood gasses. The reason for this is not because the blood gasses are not measured on time but because of the choice of sampling frequency. To deal with missing data, Cismondi et al. [55] proposed three steps: (a) a reliable method to align misaligned, unevenly sampled data using gridding and templating, (b) a statistical classification to differentiate absent values resulting from low sampling frequencies from those related to missingness mechanisms, and (c) an artificial intelligence technique to classify recoverable and not recoverable segments of missing data. The terms recoverable and not recoverable represent the missing values that must be imputed or deleted respectively.

Time series are said to be misaligned when their samples are not recorded with the sampling time [55]. Usually, the data is collected either evenly or unevenly. Both types of misaligned sampled data are presented Figure 3.

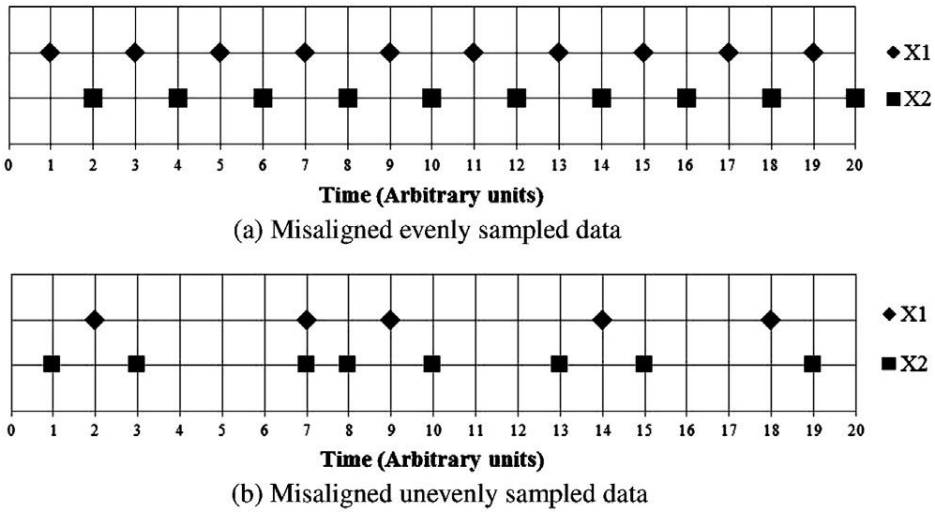

Figure 3: Simplified representation of automatically collected data [55].

Since, it is simple to align misaligned evenly sampled data, Cismondi et al. [57] suggest the methods used to deal with misaligned unevenly sampled data.

To align misaligned, unevenly sampled data, the utilization approach of two alignment methods, namely, gridding and templating, are employed. Briefly, the gridding approach consists of aligning all variables according to a fixed sampling rate. Then the variables of all samples are shifted to occupy the nodes of the grid [56]. If several data points are shifted to the same node, then cubic interpolation determines the value that is positioned on that grid [57]. At this stage of the process, these values are blank and were further be determined by applying data imputation techniques.

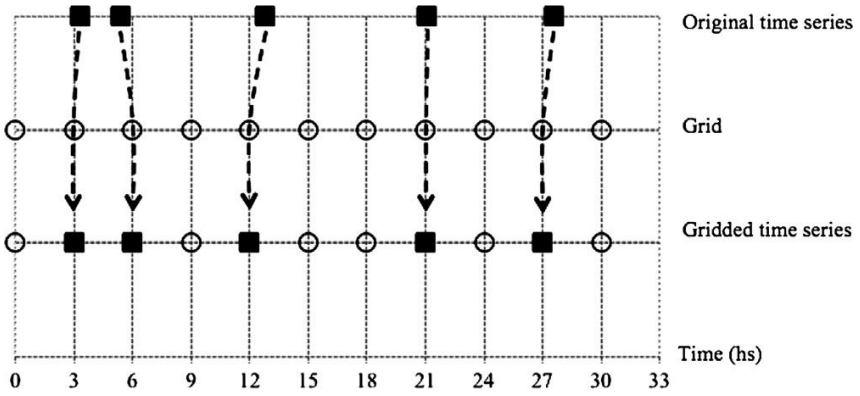

Figure 4: Representation of the gridding process [55].

Although templating is a similar technique used for correcting misaligned time series, in templating the fixed-sampling-period grid, one of the variables in the data set is replaced with the sampling times. Figure 5 represents this approach.

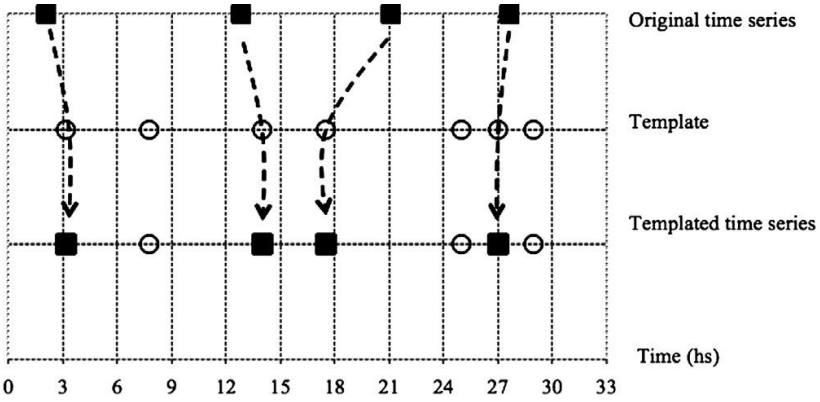

Figure 5: Representation of a template variable [55]

Usually, the most occurring variable is chosen as a templating variable.

As a second step, the imputation process of missing values is initiated. However, it is important to understand the reason for this missingness. Generally, there are two groups that explain such a reason: (1) The variable is not measured intentionally. For example, in an intensive care unit, a patient may be taken to surgery; hence, the variables are not recorded during that time. Imputing these kinds of variables would result in significantly biased predictions. Thus, this group is defined as not-recoverable missing data. (2) The variable is measured but not recorded for unidentifiable reasons. The reasons could vary from communication errors to electricity failures. This group is defined as recoverable missing data.

Next, the missing data imputation technique is applied to recover the missing value. Well known imputation techniques, such as partial imputation, partial deletion, full analysis and interpolation, are used to deal with missing values in real-world databases.

Finally, after incorporating preprocessing techniques, the data is normalized using min/max normalization. Such an approach is essential for reducing the influence of large values.

3.3 Applying Probabilistic Fuzzy Systems for Classification to Mortality Risk Prediction

This section describes the modeling method used to predict the mortality risk of septic shock patients using the Probabilistic Fuzzy Systems classifier, which was elaborated on in the previous chapter. First, the description of a problem is given. Then, the available data sets are presented. Finally, the data preprocessing steps and the experimental setup are described in detail.

3.3.1 Problem Description

The main problem tackled in this work was to develop a model for predicting the mortality risk of septic shock patients within a 72-hour time frame in an ICU setting based on updated definitions. 72-hour timeframe was chosen in accordance with Hanisch et al. [68] research, since it suggests that there is a high probability that the patient will die within three days.

In mortality prediction, the outcome of a patient is either “survival” or “deceased.” Thus, we denote the outcomes with 0 and 1, respectively, where 0 represents the survival and 1 represents the decease of a patient. Since the outcome is binary, the problem is defined as a classification problem, and Probabilistic Fuzzy Systems are employed to build this model.

3.3.2 Data Sets

Two data sets were used within this work. The prediction models for a similar data set were previously built by Fialho [14] and Balentien [63] to show the applicability of PFSC as a risk prediction model. All information regarding the patients was de-identified prior to use.

Maastricht University Medical Center+ (MUMC+) – Sepsis-2

Maastricht University Medical Center+ (Maastricht UMC+) is located in the south of the Netherlands and is the result of a merger between Maastricht University and the University Hospital Maastricht. The hospital has 715 beds, including 18 general and 9 cardiothoracic ICU beds and 6 high-dependency care unit beds, 7,593 employees, 4,685 students, and approximately 27,500 annual admissions [69]. As a collaboration project between the

Eindhoven University of Technology and the MUMC+, sepsis patients' data was collected within the Intensive Care Unit between 1/1/2013 and 1/12/2016. The data set contains 653 sepsis patients, and the identification of all septic shock patients was manually performed by dr. Ronny Schnabel. In total, 480 septic shock patients were identified with definitions accepted at International Sepsis Definitions Conference (Sepsis-2) in 2001. The final set was selected based on physiological variables and inclusion criteria (Table 6).

Maastricht University Medical Center+ (MUMC+) – Sepsis-3

The second MUMC+ data set contains 299 septic shock patients out of 653 sepsis patients identified with definitions accepted at The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) in 2016. The data, collected between 1/1/2013 and 1/12/2016, includes 27 physiological and behavioral variables.

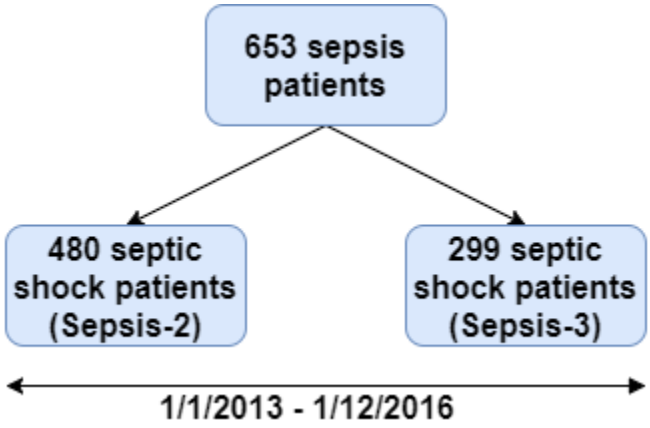

Figure 6: Graphical representation of two data sets

3.3.3 Data Preprocessing

In this subsection, the data preprocessing steps that were taken in processing the available raw data are described in detail. First, the inclusion and exclusion criteria for all septic shock patients are presented. Then, the transformation of the data format is described. After the data format, the data-cleaning technique is applied to fill the missing values, smooth the noisy data, and resolve the inconsistencies in the data. Finally, issues with features recorded in different time intervals are tackled. Following the abovementioned process, the data is ready for modeling purposes. It is important to note that since the data sets used in this work are from the same hospital, preprocessing steps, such as data integration, are deemed unnecessary.

3.3.3.1 Inclusion and Exclusion Criteria

For the MUMC+ data, patients were selected based on the inclusion and exclusion criteria. In general, inclusion criteria are characteristics that the patients must display if they are to be

included in the study. Inclusion and exclusion criteria may include factors such as age, race, sex, ethnicity, and type of disease. The general inclusion criteria for all data sets are defined as follows:

1. Age ≥ 15
2. First ICU stay
3. Existence of septic shock
4. At least one measurement for all features in Table 6 for Sepsis-2 and Sepsis-3 septic shock patients

In this project, since the prediction models are built upon on the identification of septic shock patients based on two different definitions (Sepsis-2 and Sepsis-3), one common list of features is identified and used as part of the inclusion criteria. All chosen features are supposed to be independent with minimal correlation.

Features		
Heart rate	SpO_2	Potassium
Systolic arterial blood pressure	Hematocrit	Creatinine
Diastolic arterial blood pressure	Thrombocyte	Albumin
Temperature	Leucocyte	Urea
Lactate	Mean arterial pressure	GOT (ASAT)
Breath rate	Urinary bladder catheter	GPT (ALAT)
Arterial pH	APTT	Bilirubin
Arterial pO_2	Arterial pCO_2	CRP
Bicarbonate	Sodium	Glucose

Table 6: List of features used for inclusion of Sepsis-2 and Sepsis-3 septic shock patients

Furthermore, exclusion criteria are applied to data sets to remove the outliers that are distant from other observations. Since no accurate model exists for how missing or noisy data are distributed, the interpolation of missing ICU data is extremely difficult [53]. The reason for missing data is because either the clinician believed that the data was not relevant for the observed disease and thus did not record it or different internal and external factors made the data useless [54]. Hence, values that are not available and missing for an intentional reason are removed from the data sets. The below-mentioned exclusion criteria are applied to all data sets, and the values satisfying these criteria are deleted and marked as missing values.

1. NULL, empty, or non-numerical values;
2. Wrong values that are not humanly possible (e.g., a human with -5 heart rate).
   The complete list of removal criteria is presented in Table 7.

Feature	Removal criteria
Any	NULL, empty or non-numerical value
Blood pH	< 6.5
Arterial BP	≤ 0
SpO_2	< 60%
Heart rate	≤ 0
Temperature	< 25°C and > 45°C
ALAT	> 4000
ASAT	> 5000

Table 7: List of removal criteria

3.3.3.2 Missing Values

The data collected at MUMC+ was recorded automatically using medical devices that the hospital uses. These recorded data were obtained in the format of Table 8.

PatID	Time	ItemID	Desc	Value	UoM
1	11-06-14 11:43	13	Leucocyte	2	
1	15-06-14 10:20	29	Urea	11.4	mm/L
1	15-06-14 10:20	25	Natrium	5.23	mm/L
1	15-06-14 10:20	8	Arterial pH	7.44	
1	15-06-14 10:20	9	Arterial pO_2	9.1	kPa
1	15-06-14 17:00	12	O_2 saturation	38	%

Table 8: Obtained data format from MUMC+

However, the prediction model accepts the format in Table 9.

PatID	Time	Var1	Var2	Var3	Var4
1	11-06-14 11:43	1.5	7.44	38	11.4
1	15-06-14 10:20	2	7.67	40	12
1	15-06-14 10:20	2.1	7.7	60	12.3
1	15-06-14 10:20	2.2	7.32	50	11.6
1	15-06-14 10:20	1.8	7.5	55	11.7
1	15-06-14 17:00	1.9	7.53	42	13

Table 9: Required data format from MUMC+

In order to obtain the required data format from the obtained data format, the data had to be transposed. The easiest way to achieve it could simply be transposing the column of variables into rows. However, the variables were not recorded at the same time, which would lead to empty results in most of the rows. Thus, a specific approach, described in section 3.2.3.3, was taken to resolve this issue.

3.3.3.3 Imputation of Missing Values

The data obtained from MUMC+ is unevenly sampled, thus, the utilization approach of two alignment methods, namely gridding and templating, was used in this work. The complete description of this approach is given in section 3.2.1.

Next, the imputation process of missing values was initiated. First, the mean time μ between the two samples of all features, including upper and lower bounds, was calculated and included in a confidence interval table (CI-Table). The upper and lower bounds were calculated by $\mu \pm 3 \times \sigma$. Furthermore, the second CI-Table was created, which included a patient–variable combination where each patient had $n \geq 30$ observations of that feature. Finally, the last known value was imputed if the missing value fell within either the confidence of the first CI-Table or the second CI-Table. If the missing value did not fall within this confidence bound, the whole observation was dropped.

4. Results

This chapter contains the results of the experiments. First, the results of the feature selection are presented, and the list of optimal features per data set is described. Furthermore, each model is built upon optimal features, and the corresponding results are discussed. After obtaining the results, the final models are selected and reviewed. Additionally, the performance of each model is validated using the validation set and the external data sets. Finally, the individual mortality risk estimation of an ill patient is offered and discussed.

4.1 Mortality Risk Prediction for Septic Shock Patients

Initially, the entire cohort of patients is selected while building the mortality risk prediction model for the septic shock patients. The entire cohort of patients includes the patients classified with Sepsis-2 and Sepsis-3 definitions. In order to build the final models, data preprocessing and feature selection steps are taken.

4.1.1 Data Preprocessing

Before incorporating the preprocessing steps, it is important to review the data sets and identify the impact of the cleaning process. In total, there are two data sets used for analysis purposes: MUMC+ (Sepsis-2) and MUMC+ (Sepsis-3). Each data set contains 653 sepsis patients respectively. The admission of patients with sepsis was defined by the existence of an infection and at least one organ dysfunction [59]. Using Sepsis-2 and Sepsis-3 definitions, 506 and 315 septic shock patients were identified, respectively.

After incorporating the inclusion criteria, exclusion criteria, and preprocessing steps, the final data sets are obtained. The MUMC+ (Sepsis-2) and MUMC+ (Sepsis-3) data sets contain 27 variables representing 480 and 315 septic shock patients, respectively. The descriptive statistics of both data sets are shown in Table 10.

		MUMC+ (Sepsis-2)		
		Survived	**Deceased**	**Total**
Age		62 (14.5)	64.5 (12.3)	63 (14)
Gender		308 (64.2%)	172 (35.8%)	480 (100%)
	Male	195 (63.3%)	113 (65.7%)	308 (64.2%)
	Female	113 (36.7%)	59 (34.3%)	172 (35.8%)
Weight		79.62 (22.7)	78.8 (24.4)	79.28 (23.8)
Height		172 (9)	172 (9)	172 (9)
Length of Stay		12.6 (15)	11.5 (16.8)	12.2 (15.7)
		MUMC+ (Sepsis-3)		
		Survived	**Deceased**	**Total**
Age		62.4 (14.2)	62.7 (14)	62.4 (14.1)
Gender		175 (58.5%)	124 (41.5%)	299 (100%)
	Male	113 (64.6%)	81 (65.2%)	194 (64.9%)
	Female	62 (35.4%)	43 (34.8%)	105 (35.1%)
Weight		80.82 (22.8)	80.5 (22.9)	80.47 (22.9)
Height		172.7 (9.1)	172.3 (9.4)	172.4 (9.4)
Length of Stay		12.4 (18.1)	12.4 (18)	12.4 (17.9)

Table 10: Descriptive statistics of data sets

The algorithm used to create the model accepts the data in a predefined format (Table 9) with non-empty values. The predefined format was achieved by aligning unevenly sampled data, as described in section 3.3.3.3. Since usually the most occurring variable is chosen as a templating variable, in our work, the heart rate variable is chosen for this purpose. After applying the imputation technique, observations with missing data were dropped, resulting in a reduced final data set. For the MUMC+ (Sepsis-2) data set, the initial number of observations was 81.871, and only 59.586 observations were left after the imputation, resulting in a 27.2% loss. For MUMC+ (Sepsis-3), however, the original data set contained 49.583 observations, whereas the imputed one had only 36.493 observations (-26.4%). The possible reason of this data loss is missing values that were not recorded due to various reasons (e.g. a patient may be taken to surgery). Moreover, with regard to the data loss, the possible influence may be the poor performance of the algorithm. Usually, the cause of poor performance is either overfitting or underfitting the data. Since the model is memorizing the data it has seen, it is desirable for the model to see every case and further be able to generalize to unseen examples. The data loss may result in losing the important examples, and thus, the model may not be able to capture the relationship between the input and the target values. Furthermore, for both data sets, no any patient was dropped after applying the imputation technique. The reason for such an outcome is the fact that all patients had at least one complete observation where any missing values fell within either the confidence of the first CI-Table or the second CI-Table, and thus imputed, as described in section 3.2.3.3.

Since the data sets are from the same hospital, both data sets contain the same features and the same unit of measurements. Thus, there was no mismatch among the data sets, and the data sets were ready for analysis. The mean and standard deviations of all features between the data sets are present in Table 11. These values were obtained after the alignment of the distributions.

Feature	MUMC+ (Sepsis-2)	MUMC+ (Sepsis-3)
Heart rate	94.3 (21.1)	94.3 (20.4)
Systolic arterial blood pressure	124.9 (26.5)	125.2 (26.9)
Diastolic arterial blood pressure	61.4 (13.9)	61.6 (13.7)
Lactate	3.3 (3.3)	3.9 (3.5)
Temperature	37.1 (0.9)	37.1 (0.9)
Breath rate	23.5 (7.7)	23.4 (7.6)
Arterial pH	7.4 (0.1)	7.4 (0.1)
Arterial pO_2	12.2 (3.9)	12.4 (4.0)
Bicarbonate	24.8 (6.3)	24.3 (6.6)
SpO_2	95.6 (3.2)	95.7 (3.2)
Leucocyte	13.0 (11.2)	13.2 (11.4)
Mean arterial pressure	82.7 (19.3)	83.2 (19.4)
Hematocrit	0.3 (0.1)	0.3 (0.1)
Thrombocyte	215.5 (175.1)	217.2 (183.2)
Urinary bladder catheter	159.8 (136.2)	157.4 (135.2)
APTT	46.3 (18.6)	47.4 (19.6)
Arterial pCO_2	5.5 (1.5)	5.5 (1.6)
Sodium	142.6 (7.3)	143.0 (7.1)
Potassium	4.3 (0.7)	4.3 (0.7)
Creatinine	132.7 (110.9)	129.6 (99.8)
Albumin	18.6 (7.1)	18.1 (7.4)
Urea	13.1 (8.4)	13.6 (8.6)
GOT (ASAT)	139.0 (406.7)	174.4 (480.4)
GPT (ALAT)	120.0 (312.1)	145.5 (353.3)
Bilirubin	32.2 (64.5)	31.2 (52.8)
CRP	111.9 (492.9)	139.2 (114.8)
Glucose	7.4 (2.5)	7.4 (2.5)

Table 11: Means and standard deviations (in brackets) of the features

In order to check the similarity of the data sets, one-way analysis of variance (ANOVA) was performed. The values of the same features from both data sets were analyzed separately, and the results showed that neither of the data sets with the same features have means significantly different from the other data set. For instance, the similarity analysis of the heartrate variable was performed and no significant difference was found (Figure 7).

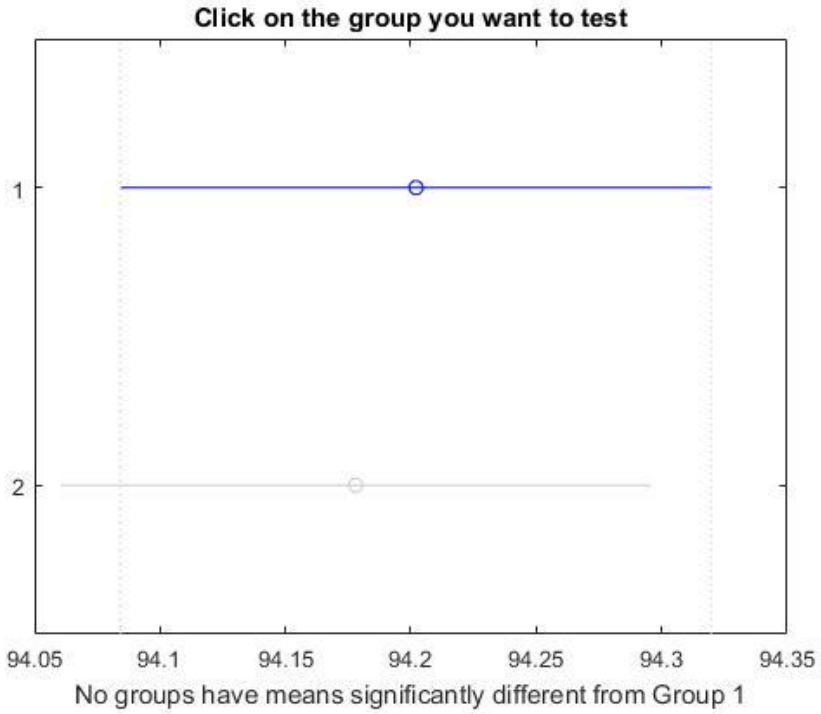

Figure 7: Similarity test of the heartrate variable

Further, the similarity test between the two data sets was performed as well. One-way ANOVA was used for this purpose and the test has not shown any significant difference between these data sets (Figure 8).

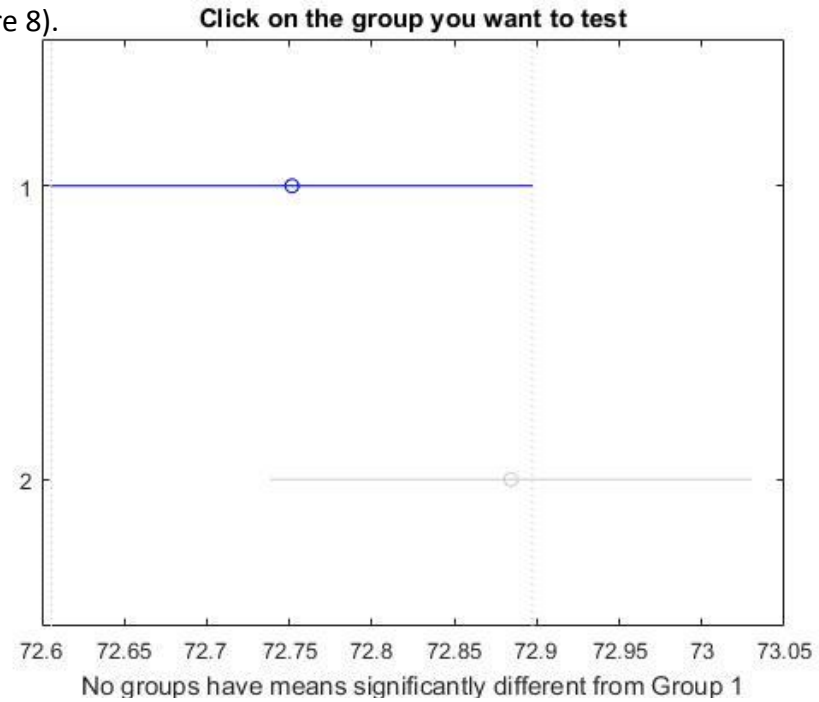

Figure 8: Similarity test of two data sets

4.1.2 Feature Selection

Following the normalization procedure, forward feature selection is employed for calculating the optimal feature set that is vital for the final model. In order to evaluate the predictive model, 10-fold cross-validation was applied, and the final optimal feature set was obtained. The complete list of features and the chosen optimal features are shown in Table 12. Since both models are from the same hospital, one would expect a bigger overlap in chosen features. However, due to the selection of septic shock patients using different definitions, namely, Sepsis-2 and Sepsis-3, only two features are shared by both models. One of the shared features is lactate, which shows a greater importance in predicting the mortality risk for septic shock patients. There have been numerous discussions on the importance of lactate in determining the severity of illness [1], [59] since that is the main addition to Sepsis-3 definitions in comparison with Sepsis-2. Also, our feature selection shows that lactate may be a new emerging vital sign of septic shock and, consequently, may be used in mortality risk prediction. Thrombocyte is another feature that is shared by both models, but further medical expert intervention is necessary for determining the reason for the significance of this variable. No other correlations between the features or across the models are found. Seventeen out of 27 features were found to be less relevant by the feature selection algorithm and thus were not included in the final model.

It is important to note that the feature selection was set to select only the top six features. The number of features lesser and greater than six was also tested, but it either had a negative effect on the accuracy of the model or did not show a significant improvement in accuracy. The order in which the features were chosen is also presented in Table 12. The blue background represents the features that are shared by both models. On the other hand, the green background, shows the features that are not shared by both models.

Variables	MUMC+ (Sepsis-2)	MUMC+ (Sepsis-3)
Heart rate		
Systolic arterial blood pressure		
Diastolic arterial blood pressure		
Lactate	2	4
Temperature		
Breath rate		
Arterial pH		1
Arterial pO_2		
Bicarbonate		
SpO_2	6	

Leucocyte		
Arterial BP MAP	5	
Hematocrit		
Thrombocyte	4	5
Urinary bladder catheter	1	
APTT [total]		6
Arterial pCO_2		
Sodium		
Potassium		
Creatinine		3
Albumin		
Urea	3	
GOT (ASAT)		
GPT (ALAT)		
Bilirubin		2
CRP		
Glucose		

Table 12: Optimal features per data set

4.1.3 Models

After obtaining the optimal feature set, the model was applied to model assessment data. The model was subjected to 10-fold cross-validation, and then the obtained model was tested on the second part, namely, the holdout set that was reserved for model validation purposes. Finally, the model was applied on other external data's holdout set to classify the data.

First, the models, the corresponding rule probabilities and parameters, and the detailed explanation of each rule are presented. The performance of each model on the other data set is then given and discussed.

4.1.3.1 MUMC+ (Sepsis-2)

As presented in Table 12, the optimal features for MUMC+ (Sepsis-2) model are Lactate, O_2 saturation, Arterial blood pressure, Thrombocytes, Urinary bladder catheter, and Urea. While training the model, Probabilistic Fuzzy Systems were set to generate two rule sets, and the corresponding probability parameters are given in Table 13.

Rule	Survived	72-hr mortality
1	2.66%	97.34%
2	97.35%	2.65%

Table 13: Probability parameters per rule for MUMC+ (Sepsis-2) model

From the Table 13, we can clearly identify two groups, where Rule 1 and Rule 2 correspond to a 97.34% mortality chance within 72 hours and a 97.35% survival chance respectively. In order to better understand these rules, PFS presents a graphical representation that is easy to interpret for a person who is not familiar with the healthcare domain.

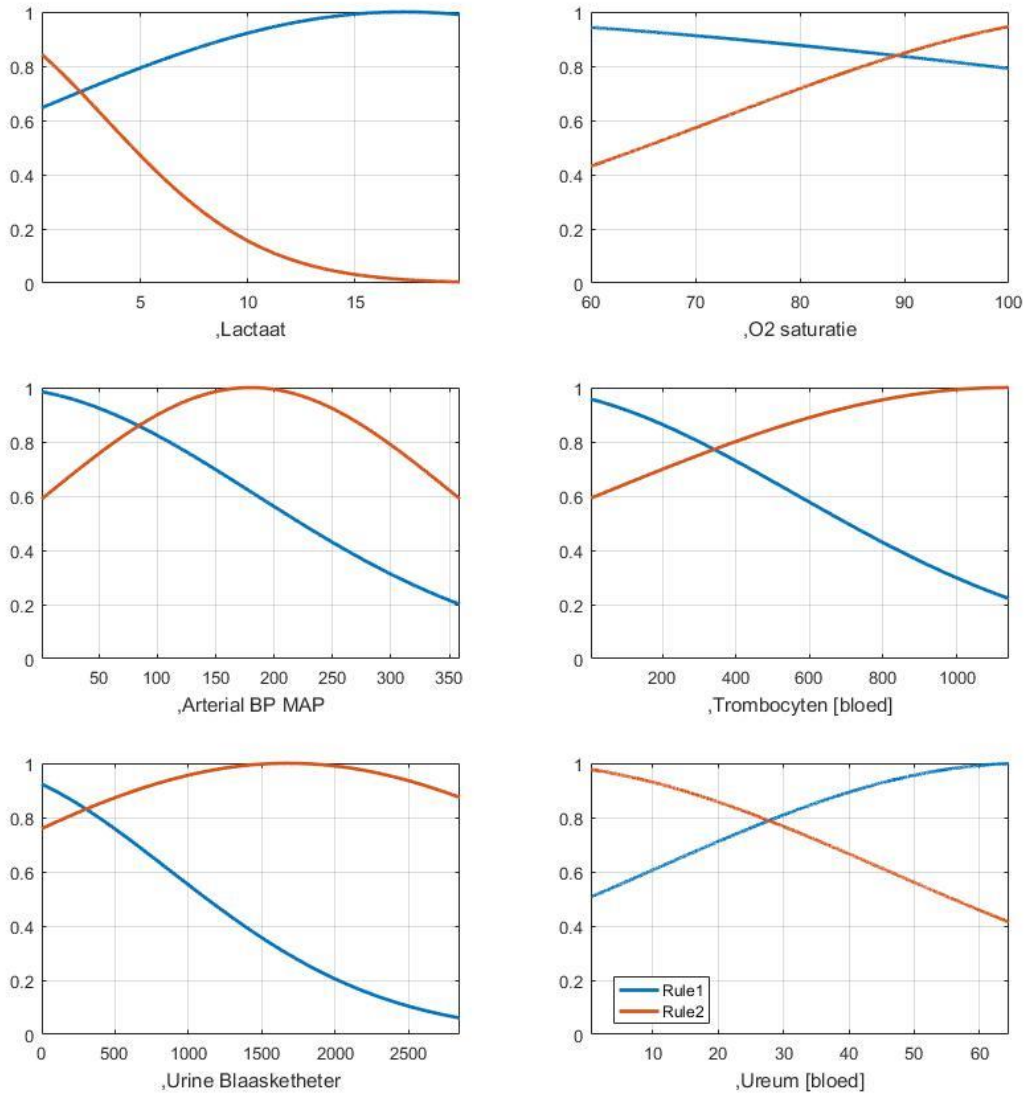

Figure 9: Graphical representation of MUMC+ (Sepsis-2) model

In Figure 9, the resultant if-then rules are the following:

1. Rule 1 (blue): **If** Lactate and SpO_2 are high, Arterial Blood Pressure is low, Thrombocytes are low, Urine Bladder Catheter is very low, and Urea is very high, **then** the patient will die within 72 hours with a probability of 97.34%.
2. Rule 2 (red): I**f** Lactate is very low, SpO_2 is very high, Arterial BP MAP is normal, Thrombocytes are very high, Urine Bladder Catheter is very high, and Urea is normal, **then** the patient will survive with a probability of 97.35%.

The other interesting observation is related to lactate. The feature selection technique chooses the most relevant variables and may reveal important but unconventional variables in the healthcare realm. With the addition of Probabilistic Fuzzy Systems, we can further help to determine the actual breakpoint between diagnostic measurements. For instance, although lactate is not a part of Sepsis-2 definitions, feature selection has considered this variable as one of the most important ones. From the graph, we can see that the lactate breakpoint is at 2 mmol/L, which actually aligns with the updated definitions of sepsis (e.g., Sepsis-3). A lactate value bigger than 2 mmol/L denotes an increased risk of mortality.

The parameters of the corresponding rules are given in Table 14.

Variable	Rule Number	v_j	σ_j
Lactate	Rule 1	0.8696	0.9089
	Rule 2	-0.1898	0.3593
SpO_2	Rule 1	0.1985	1.1709
	Rule 2	1.1381	0.4148
Arterial BP MAP	Rule 1	-0.1061	0.6171
	Rule 2	0.5015	0.4850
Thrombocytes	Rule 1	-0.1991	0.6910
	Rule 2	0.9925	0.9654
Urinary bladder catheter	Rule 1	-0.2034	0.5082
	Rule 2	0.5884	0.7925
Urea	Rule 1	1.0417	0.8864
	Rule 2	-0.1855	0.8931

Table 14: Parameters v_j and σ_j for each rule of MUMC+ (Sepsis-2) model

$\mathbf{v_j}$ and $\boldsymbol{\sigma_j}$ in Table 14 are the parameters obtained for the membership functions, while Table 13 presents the probability parameters $p_{j,k}$ associated to each rule. The plot of these membership functions is depicted in Figure 9.

4.1.3.2 MUMC+ (Sepsis-3)

The optimal feature subset of the MUMC+ (Sepsis-3) data set consists of Lactate, Arterial pH, Thrombocytes, APTT (activated partial thromboplastin time), Creatinine, and Bilirubin. The probability parameters of two predefined rules are shown in Table 15.

Rule	Survived	72-hr mortality
1	2.78%	97.22%
2	98.64%	1.36%

Table 15: Probability parameters per rule for MUMC+ (Sepsis-3) model

The graphical representation of both rules is given in Figure 10 and the linguistic interpretation is presented afterwards.

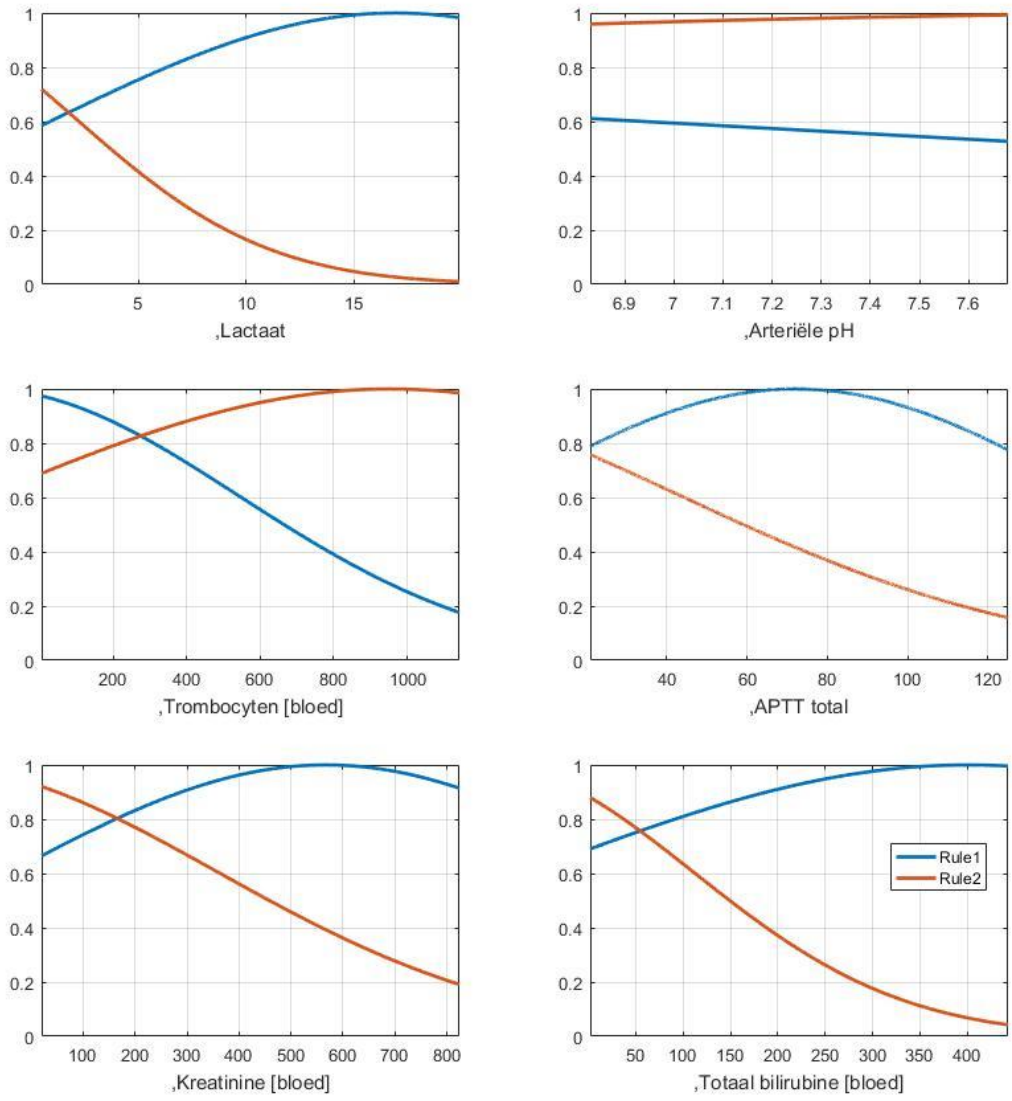

Figure 10: Graphical representation of MUMC+ (Sepsis-3) model

To better understand the rules, the Figure 10 would read as follows:

1. Rule 1 (blue): **If** Lactate is very high, Arterial pH is normal, Thrombocytes are low, and APTT, Creatinine, and Bilirubin are high, **then** the patient will die within 72 hours with a probability of 97.22%.
2. Rule 2 (red): **If** Lactate is very low, Thrombocytes are very high, and APTT, Creatinine, and Bilirubin are low, **then** the patient will survive with a probability of 98.64%.

The parameters of both rules are presented in Table 16.

Variable	Rule Number	v_j	σ_j
Lactate	Rule 1	0.8503	0.7976
	Rule 2	-0.3322	0.4398
Arterial pH	Rule 1	0.1016	0.7937
	Rule 2	1.0811	0.6638
Thrombocytes	Rule 1	-0.1352	0.6092
	Rule 2	0.8335	0.9605
APTT	Rule 1	0.5767	0.5936
	Rule 2	-0.3588	0.7065
Creatinine	Rule 1	0.6881	0.7353
	Rule 2	-0.2569	0.6911
Bilirubin	Rule 1	0.9054	1.0472
	Rule 2	-0.2470	0.4962

Table 16: Parameters v_j and σ_j for each rule of MUMC+ (Sepsis-3) model

Since both models only share two features, it is difficult to directly compare the rule parameters. However, those two shared features, namely Lactate and Thrombocytes, show similar split (first rule high, second rule low). In fact, the Sepsis-3 data set has fewer patients in total compared to the Sepsis-2 data set. However, Sepsis-3 patients are identified as much more critical than Sepsis-2 patients based on their mortality rate.

4.2 Mortality Risk Prediction for Septic Shock Patients with Three Days Stay in ICU

After the building the models to predict the mortality risk for septic shock patients for the entire cohort of patients within 72 hours, the next interesting step could be building a model for the same purposes but with different cohort of patients. This time, the length of stay (LOS) is selected as the main criterion in dividing the patients. Patients classified with Sepsis-2 and Sepsis-3 definitions, and maximum stay of three days are selected for building the final models. Selection of three days is based on previous works performed in non-cancer critically ill patients

and cancer patients with septic shock, where the organ failures during the first three days in the ICU were defined as the accurate outcome predictors [70].

4.2.1 Data Preprocessing

For this analysis, the same raw data sets are used. After incorporating the inclusion, exclusion criteria, and preprocessing steps, the final data sets are obtained. As a result, MUMC+ (Sepsis-2) LOS data set has 104 patients, whereas MUMC+ (Sepsis-3) data set LOS has 84 patients. The mortality rates for the data sets are 63% and 77% respectively.

4.2.2 Feature Selection

Following the data preprocessing steps, the feature selection algorithm is employed to select the most relevant variables for each of the data sets. The order in which the variables were chosen is presented in Table 17. The blue background represents the features that are shared by both models, whereas, the green background, depicts the features that are not shared by both models.

Variables	MUMC+ (Sepsis-2) LOS	MUMC+ (Sepsis-3) LOS
Heart rate		6
Systolic arterial blood pressure		
Diastolic arterial blood pressure		
Lactate	1	1
Temperature		
Breath rate		
Arterial pH	2	
Arterial pO_2		
Bicarbonate	4	
SpO_2	3	
Leucocyte		4
Arterial BP MAP	6	
Hematocrit		
Thrombocyte		
Urinary bladder catheter	5	
APTT [total]		3
Arterial pCO_2		
Sodium		
Potassium		

Creatinine		5
Albumin		
Urea		
GOT (ASAT)		
GPT (ALAT)		
Bilirubin		2
CRP		
Glucose		

Table 17: Optimal features per data

4.2.3 Models

After obtaining the optimal feature set, the model was applied to data that was reserved for model assessment. Next, the model was subjected to the 10-fold cross-validation. Following the cross-validation phase, the model was tested on the holdout set and on external data's holdout set consecutively. Finally, the corresponding rule probabilities and parameters, and the performance of each model on the other data set were obtained

4.2.3.1 MUMC+ (Sepsis-2) LOS

In order to generate the rules of the MUMC+ (Sepsis-2) LOS model, optimal features are required. As depicted in Table 17, the following features are selected from the FS data subset: Lactate, Arterial pH, Bicarbonate, SpO_2, Arterial BP MAP, Urinary bladder catheter. Two rulesets and corresponding probability parameters are given in Table 18.

Rule	Survived	72hr mortality
1	97.5%	2.5%
2	2.3%	97.7%

Table 18: Probability parameters per rule for MUMC+ (Sepsis-2) LOS model

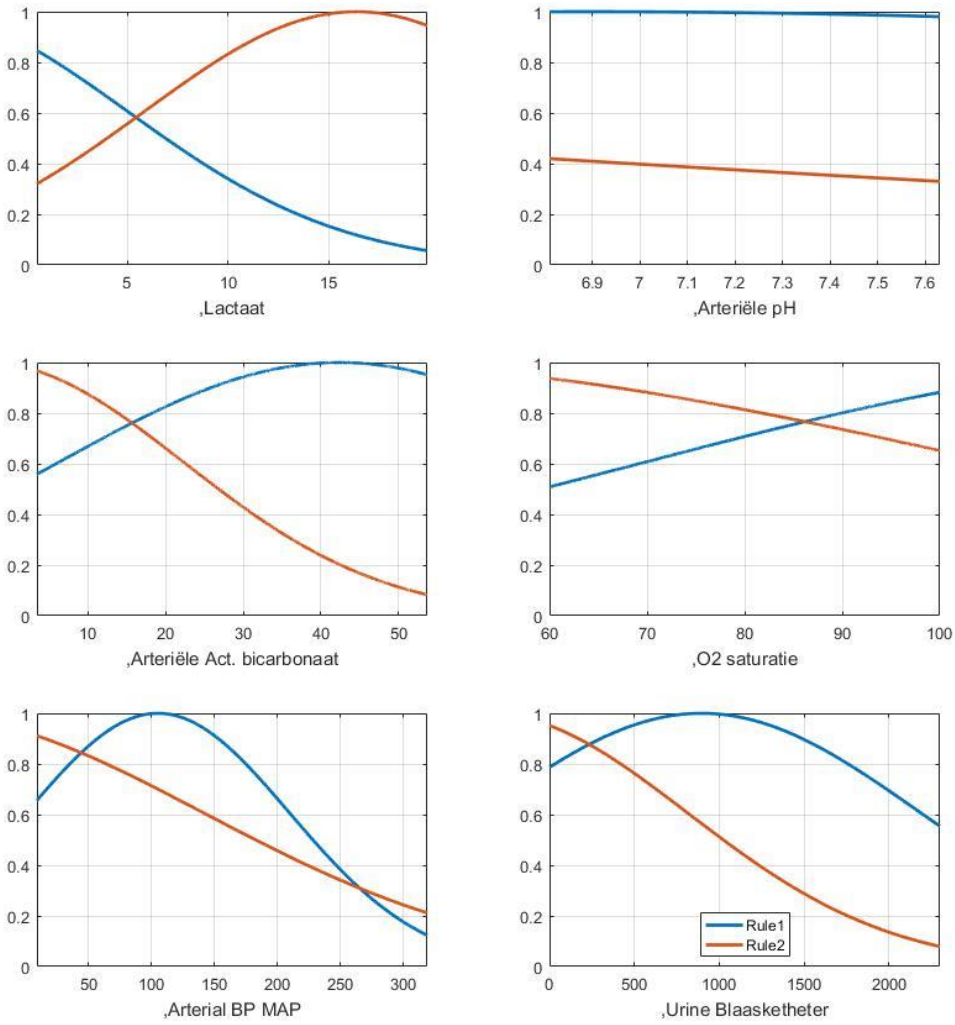

Figure 11: Graphical representation of MUMC+ (Sepsis-2) LOS model

The linguistic interpretation of the rules given in Figure 11 would read as follows:

1. Rule 1 (blue): **If** Lactate is very low, Bicarbonate are very high, SpO_2 is normal, Arterial BP MAP is very low, and Urinary bladder catheter is very low, **then** the patient will survive within 72 hours with a probability of 97.5%.
2. Rule 2 (red): **If** Lactate is very high, Bicarbonate are very low, SpO_2 is high, Arterial BP MAP is very low, and Urinary bladder catheter is normal, **then** the patient will die within 72 hours with a probability of 97.7%.

As previously built models (e.g. MUMC+ (Sepsis-2), MUMC+ (Sepsis-3), there is a clear separation between two groups, one where a patient survives with 97.5% probability or dies with a probability of 97.5% corresponding to rule 1 and rule 2 respectively. Four features, namely Lactate, SpO_2, Arterial BP MAP and Urinary bladder catheter overlaps with the features of MUMC+ (Sepsis-2) model where the mortality risk was predicted for the entire cohort of patient.

Parameters v_j and σ_j for each rule of MUMC+ (Sepsis-2) LOS model for each rule are given in Table 19.

Variable	Rule Number	v_j	σ_j
Lactate	Rule 1	-0.2842	0.5356
	Rule 2	0.8224	0.5284
Arterial pH	Rule 1	0.9031	0.4863
	Rule 2	0.0669	0.6265
Bicarbonate	Rule 1	0.7885	0.6724
	Rule 2	-0.0596	0.4750
SpO_2	Rule 1	1.3030	0.6049
	Rule 2	0.3449	0.7083
Arterial BP MAP	Rule 1	0.3292	0.3280
	Rule 2	-0.2864	0.7298
Urinary bladder catheter	Rule 1	0.3893	0.5622
	Rule 2	-0.1627	0.5165

Table 19: Parameters v_j and σ_j for each rule of MUMC+ (Sepsis-2) LOS model

4.2.3.2 MUMC+ (Sepsis-3) LOS

The optimal feature set for the MUMC+ (Sepsis-3) model consists of Heart Rate, Lactate, Leucocyte, APTT, Creatinine and Bilirubin. The probability parameters of the obtained two rules are shown in Table 20. Yet again, there is a clear separation between two classes.

Rule	Survived	72hr mortality
1	98.6%	1.4%
2	1.6%	98.4%

Table 20: Probability parameters per rule for MUMC+ (Sepsis-3) LOS model

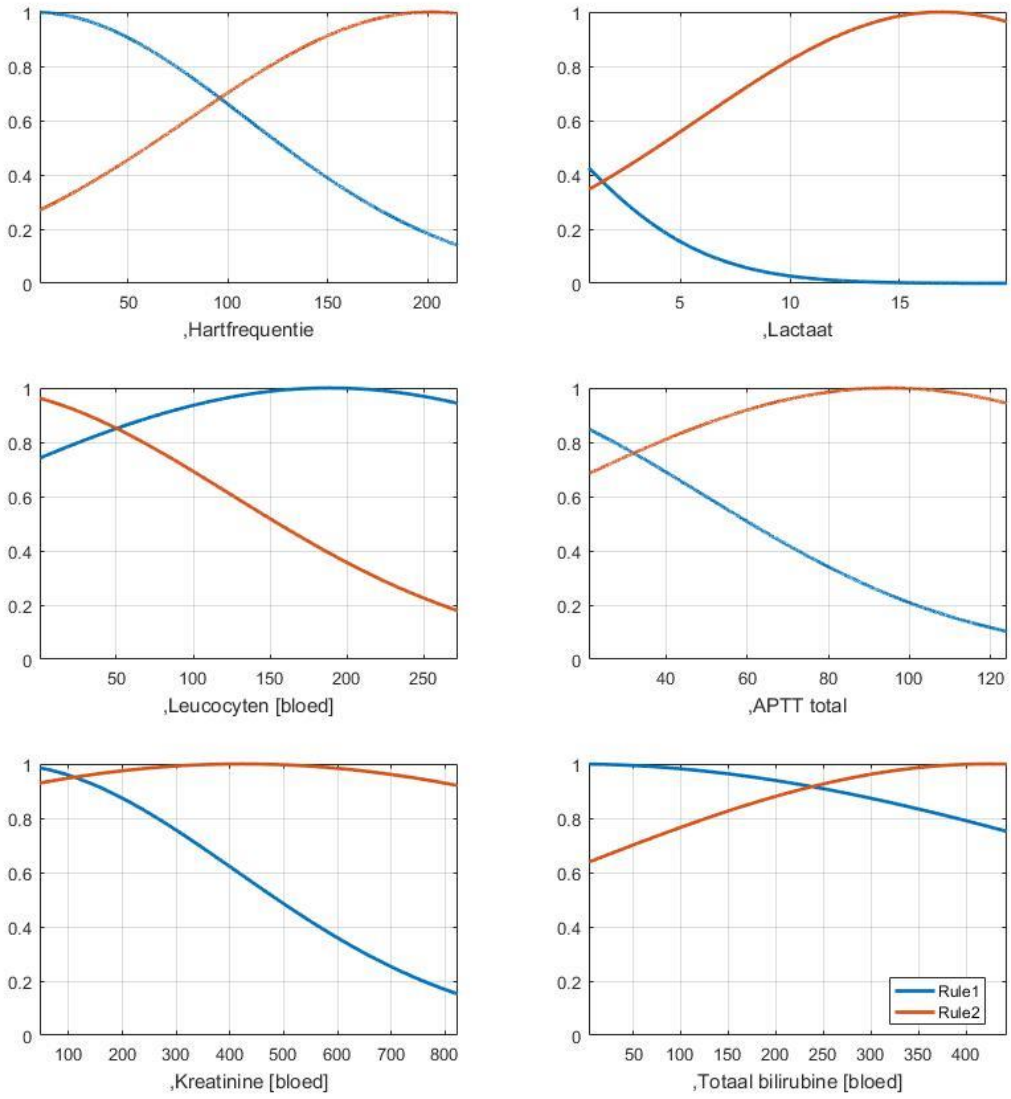

Figure 12: Graphical representation of MUMC+ (Sepsis-3) LOS model

Rules, depicted in Figure 12, can be interpreted as follows:

1. Rule 1 (blue): **If** Heart Rate, Lactate, APTT and Creatinine are very low, Leucocyte and Bilirubin are very high, **then** the patient will survive within 72 hours with a probability of 98.6%.
2. Rule 2 (red): **If** Lactate is very high, Bicarbonate are very low, SpO_2 is high, Arterial BP MAP is very low, and Urinary bladder catheter is normal, **then** the patient will die within 72 hours with a probability of 98.4%.

Lastly, the parameters of each rule of MUMC+ (Sepsis-3) LOS model are given in Table 21.

Variable	Rule Number	v_j	σ_j
Lactate	Rule 1	-0.2842	0.5356
	Rule 2	0.8224	0.5284
Arterial pH	Rule 1	0.9031	0.4863
	Rule 2	0.0669	0.6265
Bicarbonate	Rule 1	0.7885	0.6724
	Rule 2	-0.0596	0.4750
SpO_2	Rule 1	1.3030	0.6049
	Rule 2	0.3449	0.7083
Arterial BP MAP	Rule 1	0.3292	0.3280
	Rule 2	-0.2864	0.7298
Urinary bladder catheter	Rule 1	0.3893	0.5622
	Rule 2	-0.1627	0.5165

Table 21: Parameters v_j and σ_j for each rule of MUMC+ (Sepsis-3) LOS model

4.3 Model Performance

In this subsection, the performance of the models generated with Probabilistic Fuzzy Systems is analyzed. Our focus is on the analysis of each model performed within internal cross-validation, on the holdout sample, and on other external data sets. Area Under Curve (AUC), Sensitivity, and Specificity are selected as the performance metrics, and all associated results are presented in Table 22.

Metric	Model	CV	MUMC+ (Sepsis-2)	MUMC+ (Sepsis-3)	MUMC+ (Sepsis-2) LOS	MUMC+ (Sepsis-3) LOS
AUC	MUMC+ (Sepsis-2)	0.73 (0.02)	0.75	0.71	0.68	0.72
	MUMC+ (Sepsis-3)	0.77 (0.02)	0.70	0.65	0.62	0.69
	MUMC+ (Sepsis-2) LOS	0.89 (0.02)	0.82	0.85	0.83	0.83
	MUMC+ (Sepsis-3) LOS	0.89 (0.02)	0.85	0.83	0.86	0.74
Sensitivity	MUMC+ (Sepsis-2)	0.65 (0.04)	0.67	0.57	0.26	0.49
	MUMC+ (Sepsis-3)	0.76 (0.02)	0.57	0.53	0.18	0.50
	MUMC+ (Sepsis-2) LOS	0.90 (0.04)	0.79	0.80	0.54	0.81
	MUMC+ (Sepsis-3) LOS	0.85 (0.05)	0.67	0.93	0.32	0.81
Specificity	MUMC+ (Sepsis-2)	0.81 (0.03)	0.72	0.77	0.98	0.85
	MUMC+ (Sepsis-3)	0.78 (0.04)	0.76	0.81	0.99	0.85
	MUMC+ (Sepsis-2) LOS	0.89 (0.05)	0.67	0.87	0.95	0.61
	MUMC+ (Sepsis-3) LOS	0.93 (0.02)	0.87	0.57	0.91	0.69

Table 22: Performance metrics of the models on the other data sets

From Table 22, we can observe that the performance of the MUMC+ (Sepsis-2) model is slightly better than the performance of other models, but such a small difference makes these models comparable. It is also important to note that the performance of almost all models is slightly

better than the work of Balentien [63]. The main reason could be the quality and size of the actual data. MUMC+ (Sepsis-2) LOS and MUMC+ (Sepsis-3) LOS models have shown good discriminatory power (AUC>0.80). Surprisingly, the model trained on the smallest data set, namely MUMC+ (Sepsis-3) LOS model, performed satisfactory.

Due to the fact that septic shock is considered a lethal syndrome, high sensitivity is preferred over specificity. However, our tool is focused on predicting the mortality rather than the existence of the syndrome. Thus, in our case, high specificity is preferred over high sensitivity. Further, since high specificity represents the proportion of patients expected to be deceased, the urgent alignment of treatment may be employed to save the life of a patient. Derived from this discussion, we can investigate each value of the models.

Table 22 shows that the both models (MUMC+ (Sepsis-2) and MUMC+ (Sepsis-3) have demonstrated high specificity results, whereas the sensitivity of both the MUMC+ (Sepsis-2) and MUMC+ (Sepsis-3) models are lower in other data sets. Certainly, having equally high results on both performance metrics would be ideal. However, such results are acceptable since we are more interested in correctly predicting the truly ill patients and providing the necessary and immediate care needed to save those patients.

Yet another interesting results were obtained after applying the models built upon the old definitions (MUMC+ (Sepsis-2) and MUMC+ (Sepsis-2) LOS) to models built with updated definitions (MUMC+ (Sepsis-3) and MUMC+ (Sepsis-3) LOS). Table 22 shows that the MUMC+ (Sepsis-2) LOS model performs better on almost all data sets. In general, the models satisfying the old definitions have demonstrated better results than the models based on updated definitions. One of the possible reasons could be the bigger sample size used for training of these models with old definitions. Therefore, according to obtained results, a model trained on old definitions can be used on new ones. However, if we want to see the applicability of our old model on the future data sets that may be collected using the updated definitions, then it is important to look at the nature of these new definitions. If the new definitions consist of the variables used for building the old model, then the old model may perform satisfactory in predicting mortality risk of septic shock patients classified with new definitions (Figure 13). On the contrary, if the future definitions incorporate any variables that have never been used in building the old model, then our model cannot be used, and thus, the new model has to be retrained upon new definitions (Figure 14).

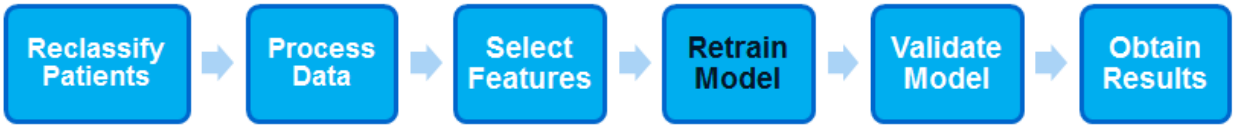

Figure 13: Updated definitions incorporate previously used variables

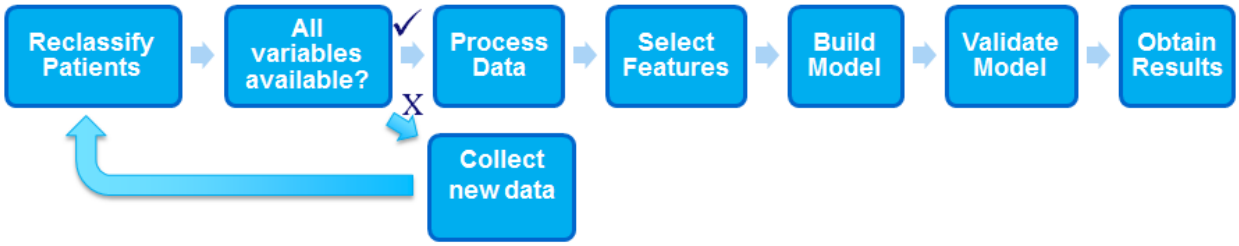

Figure 14: Updated definitions incorporate new variables

The Area Under the Curves of the respective models depicted in Table 22 are graphically represented in Figure 15, Figure 16, Figure 17 and Figure 18. ROC curves depicted in all figures, helps us to discard the suboptimal models and select the optimal models. ROC graphs in Figure 15 and Figure 16 indicate that the MUMC+ (Sepsis-2) model performs slightly better on the entire cohort of patient, whereas MUMC+ (Sepsis-2) LOS model performs better on the cohort of patients with at most 3 ICU days. Since the model generalizes better with bigger training set size, one of the possible reasons for both cases could be the size of the respective training sets in comparison with the other data sets' training sets.

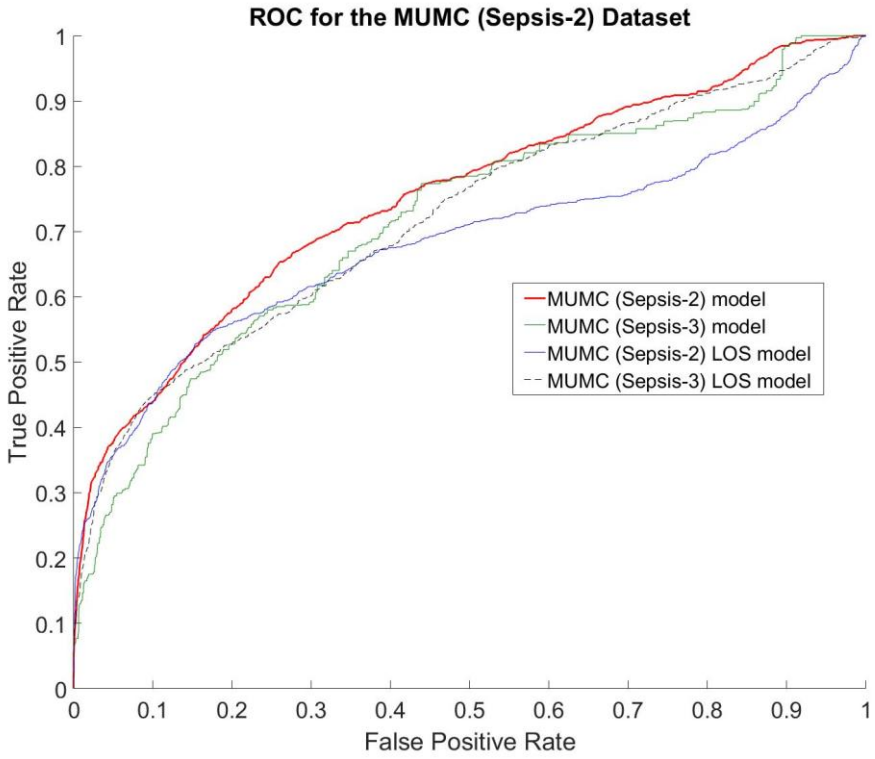

Figure 15: ROC curves of all models for the MUMC+ (Sepsis-2) data set

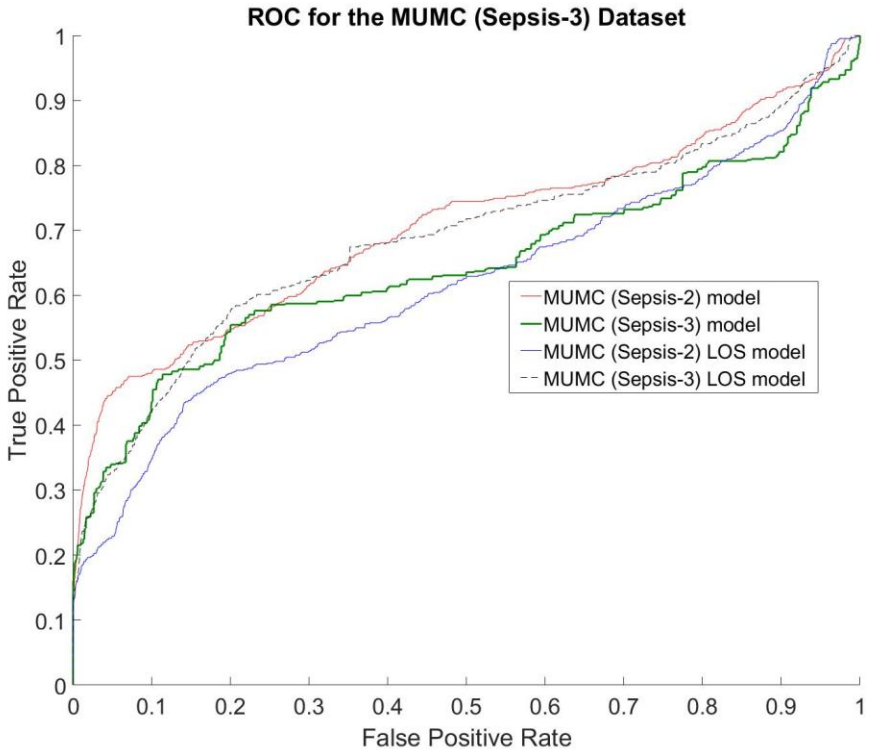

Figure 16: ROC curves of all models for the MUMC+ (Sepsis-3) data set

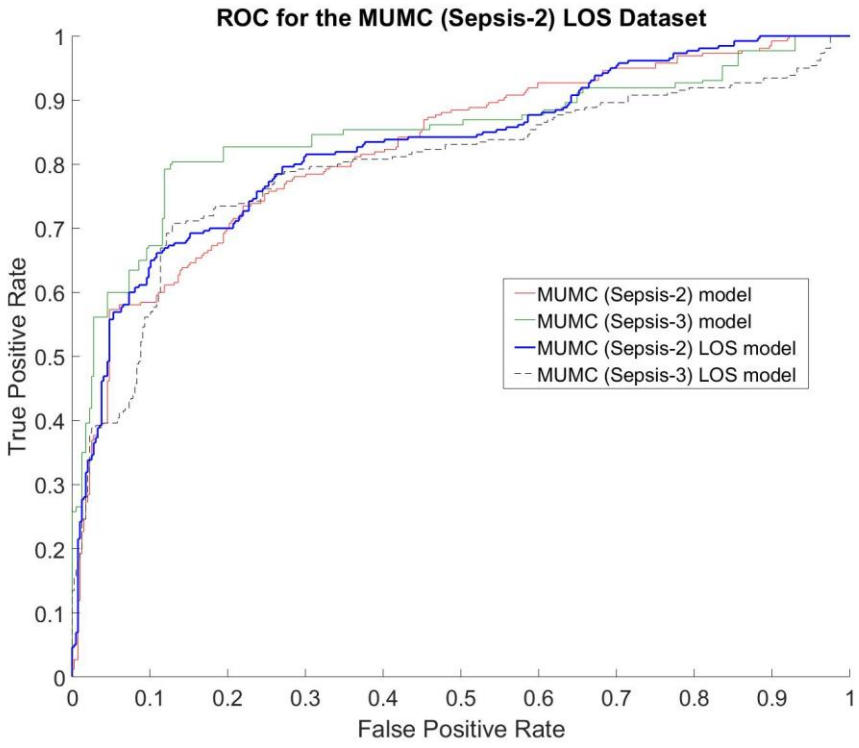

Figure 17: ROC curves of all models for the MUMC+ (Sepsis-2) LOS data set

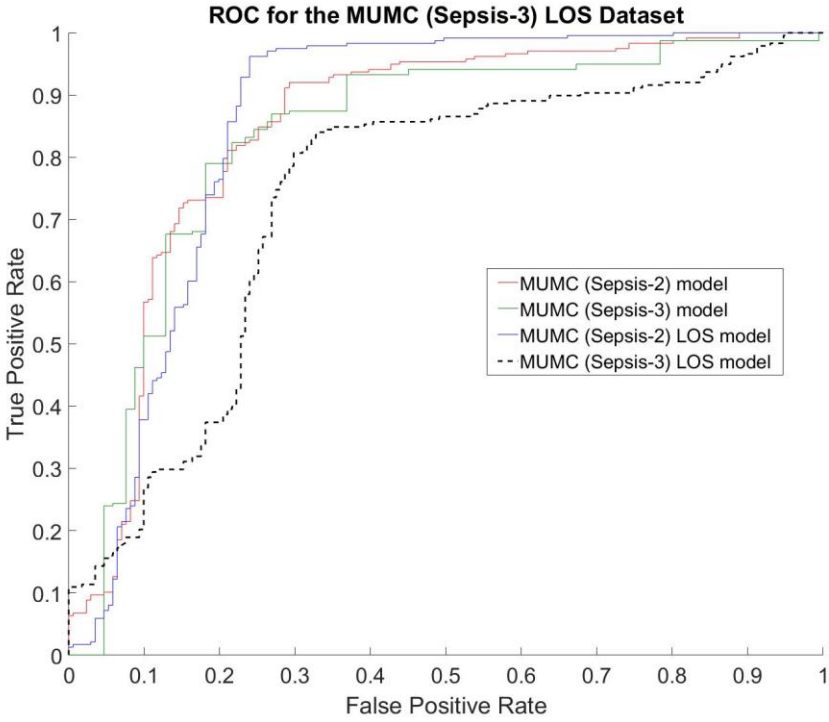

Figure 18: ROC curves of all models for the MUMC+ (Sepsis-3) LOS data set

Lastly, it is important to consider the effort required for building the models based on the updated definitions. At the first stage, four months were spent on acquiring the initial data. Then it took three days to reclassify the patients using the old and updated definitions. At the next stage, one week was spent on formatting the data that model requires. Finally, two hours were spent on training the MUMC+ (Sepsis-2) model and one hour on the other models.

4.4 Individual Mortality Risk Estimation

The final model is trained on the cohort of patients admitted to the ICU between 1/1/2013 and 31/12/2016. The predictive power of our model allows physicians to also apply the prediction to newly admitted patients. Based on the newly obtained values of a patient, the model may help doctors identify the severity of the illness and thus align the type of treatment for that specific patient.

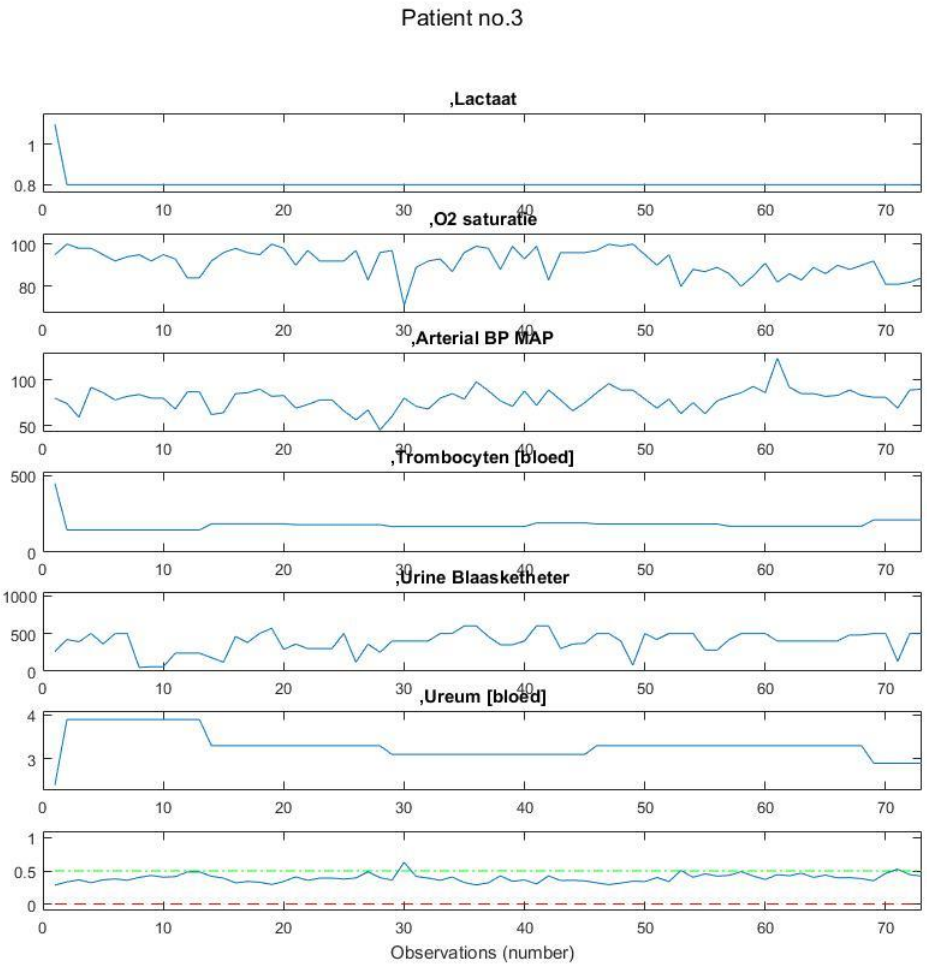

Figure 19: Mortality risk estimation of patient no. 3

In Figure 19, the Y-axis denotes the actual measurement of a patient's physiological variable, whereas the X-axis represents the total number of observations. The blue line in the last row represents the mortality risk estimation, whereas the red line denotes the actual survival of a patient. Due to the fact that our observations were templated with the heart rate, and since the heart rate was measured every 2 hours, every 12 observations are equal to 24 hours. Thus, in order to estimate the mortality risk of a newly admitted septic shock patient, the blue line in the last row of Figure 19 should be considered. The model is not limited to a particular number of hours, and depending on the measurements, the particular timeframe may be adjusted within the model and the severity of illness of a patient may be obtained. For instance, although this patient has survived, in 30th observation (approximately at the end of the third day) the mortality risk of a patient was higher. Most probably, since the clinicians intensified the treatment of this particular patient, the mortality risk went down and thus saved the life of an ill patient.

Figure 20 depicts the mortality risk estimation of another patient. From the last row of the graph, we can observe that the mortality risk estimator was constantly changing throughout

the number of observations. For example the mortality risk of a patient between 300 and 400 observations (between the twenty-fifth and thirty-third days) were estimated as high. After the intervention of the clinician by increasing the intensity of a treatment, the patient's mortality risk went down and the patient survived.

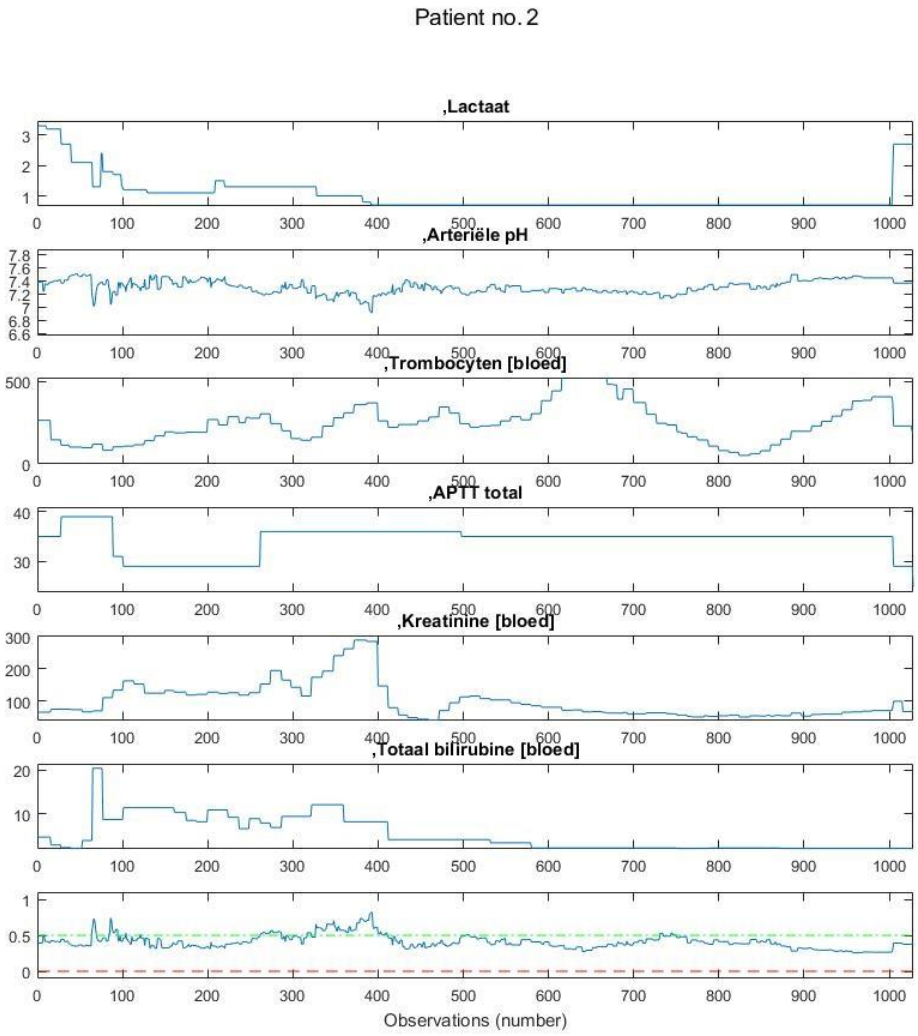

Figure 20: Mortality risk estimation of patient no. 2

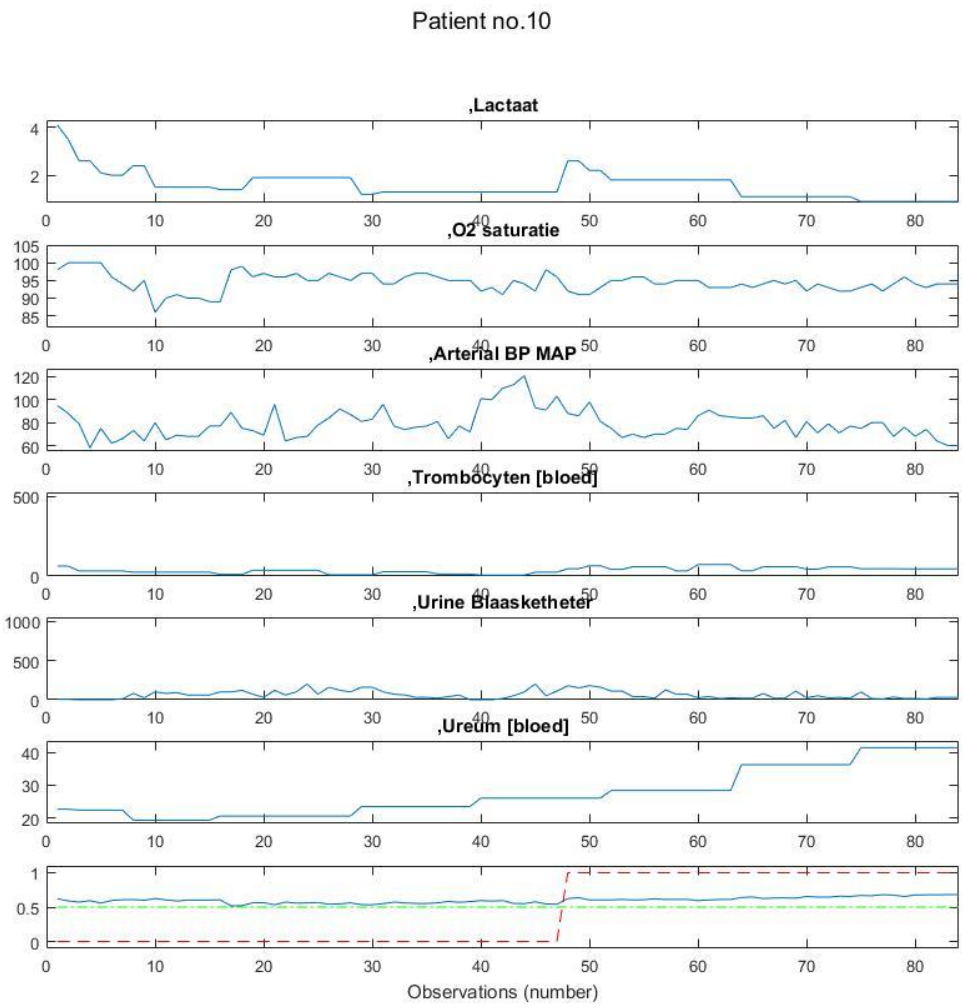

Figure 21: Mortality risk estimation of patient no. 10

Lastly, the Figure 21 depicts a mortality risk estimation of another patient who was actually deceased. Mortality risk estimator, represented by the blue line, has correctly raised an alarm about the patient's condition and eventually the patient was deceased.

5. Conclusions

This study was set out to identify the model migration approach due to change of definitions. The analysis was conducted using Probabilistic Fuzzy Systems for Mortality Risk Prediction based on old and updated sepsis definitions. Before discussing the major findings and contributions, it is appropriate to restate our main research question, which is defined as follows:

Research Question: *How to migrate the model due to change of definitions and what are the consequences of that change on the modeling activities?*

Prior to investigating the problem in detail, we discussed the applicability of scoring models in identifying the mortality of septic shock patients and indicated the advantages of prediction models in comparison to those traditional methodologies. We then introduced Probabilistic Fuzzy Systems as being a crucial approach toward modeling vagueness in linguistic terms and probabilistic uncertainty simultaneously. This approach provides doctors with an easier interpretation of the obtained results from prediction models.

The main contribution we made with this work is a developed approach toward the migration of models when the definitions of a disease change, following the description of consequences of these changes on the modeling activities. For instance, if the new definitions incorporate the variables that were used to build the previous models, then the existing models can safely be migrated and used for mortality prediction by simply re-identifying the patients. If, however, definitions get updated with new variables that have never been used for building the models, then the re-identification of patients is impossible; thus, the hospital has to restructure its data-recording mechanism and conduct the complete analysis using the new data.

The next contribution was a data analysis of septic shock patients' mortality using Probabilistic Fuzzy Systems built upon updated septic shock definitions. We have shown that PFS can perform satisfactorily for such analysis, regardless of the definition update. After predicting the mortality of patients classified with Sepsis-2 and Sepsis-3 definitions, the obtained results were compared, and the relevance of the updated definitions was identified. In this work, we have shown that the Sepsis-3 definitions demonstrate comparable results with Sepsis-2 definitions and that the identification of the presence of septic shock in critically ill patients may be performed using updated definitions. Undoubtedly, such a conclusion is drawn purely based on mathematical analysis and may be different from medical analysis. While comparing the performance of the models, however, it is essential to discuss the size of the patient cohort and the respective severity of illness. For physicians, critically ill patients are of the utmost priority since such patients need an urgent intervention that could improve their outcome. For this

reason, Sepsis-3 definitions are favored over Sepsis-2 definitions. Sepsis-3 definitions consist of the desired cohort of patients with higher mortality risk (41.5% vs. 35.8% for the entire cohort and 77% vs. 63% for the "at most three days stay in ICU" cohort) and save resources and time for physicians by focusing on the treatment of severely ill patients. Nevertheless, in order to identify patients with the updated definitions, enormous efforts must be spent in changing the structure of the hospitals. It is believed that while the practice of identification with Sepsis-2 definitions has gained some traction in healthcare, changing these definitions could "set back decades of work and could be interpreted to de-emphasize intervention at earlier stages when the syndrome is most treatable" [61].

A further contribution is obtaining individual mortality risk estimation for septic shock patients. This method may help doctors identify the severity of the illness and align the type of treatment for that ill patient accordingly, which has a direct impact on a patient's survival. Since the early detection of illness in septic shock patients results in lower mortality [16], the developed prediction tool may assist clinicians in this regard.

To conclude, in this project, we have analyzed and presented a model for predicting the mortality risk of critically ill patients identified with Sepsis-3 definitions, which constitutes a relatively new research approach for the medical field. We also compared the results with the mortality risk of Sepsis-2 patients and presented the relevance of new definitions in a mathematical way. We believe that the developed model generates reliable results to help clinicians select the appropriate treatment for each individual patient. It is important to note, however, that this tool cannot be used alone and must be treated as a complementary tool to a well-developed clinical assessment and other laboratory results.

5.1 Limitations and Future Work

In this section, the observed limitations of our implemented framework and the possibilities for future research within the septic shock mortality risk prediction domain are presented. Four major directions are identified and presented below.

5.1.1 Data Quality

The first limitation is related to the quality of data. During the data preprocessing steps, the missing values were analyzed and imputed if the data was not recorded for unidentifiable reasons, such as electricity or hardware failure. However, this approach adds bias to the prediction results, whereas deleting the missing values causes both bias and a loss of statistical power [55]. Thus, while collecting the data, it is important to prevent circumstances that affect the number of missing values.

5.1.2 Data Relevancy

The second limitation of our method is related to the relevancy of the data. The hospital, particularly Maastricht University Medical Center+, was not recording the patient data explicitly for mortality risk prediction but as a general practice within the Intensive Care Unit. Hence, collecting the specific data, which includes all relevant variables of the updated definitions, is of the utmost importance. After that, the direct comparison of models built upon Sepsis-2 and Sepsis-3 definitions may be carried out, and the most promising definition in terms of results may be employed for the identification of septic shock patients.

5.1.3 Medical Expert

The third important aspect of future work is related to the study of prepared models by a medical expert. Due to the fact that our analysis was conducted outside of a medical environment, the crucial constant feedback sessions with the doctors were not possible for numerous reasons. Such analysis, however, requires extensive input by professional physicians in order to obtain the most accurate and relevant prediction results. For example, during the data-cleaning process, six variables were removed from the analysis due to many missing values of patients. The importance of those variables was not discussed with medical experts, and thus, in the future, it is essential that this step is made obligatory in order to ensure the alignment between data analysis professionals and healthcare experts.

5.1.4 The Diagnosis of Sepsis

Since the introduction of Sepsis-3 definitions, numerous discussions have circulated within the sepsis domain. The updated definitions were introduced by the task force convened by the European Society of Intensive Care Medicine and the Society of Critical Care Medicine. The task force recommendations and the supporting evidence were endorsed by 31 medical societies. However, this document was not endorsed by all medical societies (i.e., Latin American Sepsis Institute, American College of Chest Physicians and American College of Emergency Physicians). According to these opposing societies, newly introduced qSOFA and SOFA criteria are not the actual diagnosis of sepsis but rather predictors of mortality. Furthermore, there is still no known precise pathophysiological feature that defines sepsis [62], thus discrediting the validity of these criteria. In our project, however, we are interested in predicting the mortality risk of septic shock patients rather than the mortality risk of sepsis patients. Thus, for the MUMC+ (Sepsis-3) data set, the diagnosis of sepsis was carried out using Sepsis-2 definitions, namely, SIRS criteria. Later, the serum lactate level was used to select the cohort of septic shock patients according to the updated definitions. This approach was taken in consultation with Dr. Ronny Schnabel, an anesthetist-intensivist within the ICU of MUMC+ hospital since not all

variables of newly adopted SOFA were available for the diagnosis of sepsis at the time. Therefore, future work should focus on the reclassification of patients using the SOFA score and serum lactate to identify sepsis and septic shock patients respectively. Following the identification of septic shock patients, the applicability of newly established definitions may be analyzed.

6 Bibliography

[1] Shankar-Hari, Manu, Gary S. Phillips, Mitchell L. Levy, Christopher W. Seymour, Vincent X. Liu, Clifford S. Deutschman, Derek C. Angus, Gordon D. Rubenfeld, and Mervyn Singer. "Developing a New Definition and Assessing New Clinical Criteria for Septic Shock." Jama 315.8 (2016): 775.

[2] Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, Steingrub JS, Garber GE, Helterbrand JD, Ely EW, Fisher CJ Jr, Recombinant human protein C Worldwide Evaluation in Severe Sepsis (PROWESS) study group: Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med 2001, 344:699-709.

[3] Engel C, Brunkhorst FM, Bone HG, et al. Epidemiology of sepsis in Germany: results from a national prospective multicenter study. Intensive Care Med, 2007;33:606-618.

[4] Marshall JC, Vincent JL, Guyatt G, et al. Outcome measures for clinical research in sepsis: a report of the 2nd Cambridge Colloquium of the International Sepsis Forum. Crit Care Med, 2005;33:1708-1716.

[5] Kissoon N, Carcillo JA, Espinosa V, Argent A, Devictor D, Madden M, Singhi S, van der Voort E, Latour J: World Federation of Pediatric Intensive Care and Critical Care Societies: Global Sepsis Initiative. Pediatr Crit Care Med, 2011.

[6] Thomas, Benjamin S., S. Reza Jafarzadeh, David K. Warren, Sandra McCormick, Victoria J. Fraser, and Jonas Marschall. "Temporal Trends in the Systemic Inflammatory Response Syndrome, Sepsis, and Medical Coding of Sepsis." BMC Anesthesiology. BioMed Central, 24 Nov. 2015.

[7] Kaukonen KM, Bailey M, Suzuki S, Pilcher D, Bellomo R: Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000-2012. JAMA 2014, 311(13): 1308-1316.

[8] Shen HN, Lu CL, Yang HH: Epidemiologic trend of severe sepsis in Taiwan from 1997 through 2006. Chest 2010, 138(2): 298-304.

[9] Martin, Greg S., David M. Mannino, Stephanie Eaton, and Marc Moss. "The Epidemiology of Sepsis in the United States from 1979 through 2000." New England Journal of Medicine 348.16 (2003): 1546-554.

[10] Annane D, Aegerter P, Jars-Guincestre MC, et al. Current epidemiology of septic shock: the CUB-Rea Network. Am J Respir Crit Care Med, 2003; 168:165-172.

[11] Torio CM (AHRQ), Andrews RM (AHRQ). National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011. HCUP Statistical Brief #160. August 2013. Agency for Healthcare Research and Quality, Rockville, MD.

[12] Peake SL, Delaney A, Bailey M, Bellomo R, Cameron PA, Cooper DJ, Higgins AM, Holdgate A, Howe BD, Webb SA et al: The ARISE investigators. Goal-directed resuscitation for patients with early septic shock. N Engl J Med 2014, 371(16):1496-1506.

[13] Yealy DM, Kellum JA, Huang DT, Barnato AE, Weissfeld LA, Pike F, Terndrup T, Wang HE, Hou PC, LoVecchio F et al: The PROCESS investigators. A randomized trial of protocol-based care for early septic shock. N Engl J Med 2014, 370(18):1683-1693.

[14] Fialho, A. S., Kaymak, U., Almeida, R. J., Cismondi, F., Vieira, S. M., Reti, S. R., Sousa, J. M. C., and Finkelstein, S. N. (2012). Probabilistic fuzzy prediction of mortality in intensive care units. 2012 IEEE International Conference on Fuzzy Systems, (i):1-8.

[15] Hotchkiss RS, Monneret G, Payen D. Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. Nat Rev Immunol. 2013;13(12):862-874.

[16] Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, Sevransky JE, Sprung CL, Douglas IS, Jaeschke R, Osborn TM, Nunnally ME, Townsend SR, Reinhart K, Kleinpell RM, Angus DC, Deutschman CS, Machado FR, Rubenfeld GD, Webb SA, Beale RJ, Vincent JL, Moreno R. Sepsis Campaign: International Guidelines for Management of Severe Sepsis and Septic Shock: 2012. Critical Care Medicine 41(2):580–637.

[17] Sinclair WJ. Semmelweis: his life and his doctrine. Manchester: Manchester University Press, 1909.

[18] Raju TN. Ignac Semmelweis and the etiology of fetal and neonatal sepsis. J Perinatol 1999; 19(4): 307-310.

[19] Moreno R, Jordan B, Metnitz P: The changing prognostic determinants in the critically ill patient. In 2007 Yearbook of Intensive care and Emergency.

[20] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE: APACHE acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981, 9:591-597.

[21] Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A, Harrell FE: The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991, 100:1619-1636.

[22] Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA: Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Crit Care Med 1998, 26:1317-1326.

[23] Zimmerman, JE, Kramer, AA, Mcnair, DS, and Malila, FM (2006). Acute physiology and chronic health evaluation (apache) IV: hospital mortality assessment for today's critically ill patients. Critical care medicine, 34(5):1297-310.

[24] Ayazoglu, TA (2011). A comparison of APACHE II and APACHE IV scoring systems in predicting outcome in patients admitted with stroke to an intensive care unit. Anaesth, Pain & Intensive Care 2011; 15(1).

[25] Saleh, Abdelbaset, Magda Ahmed, and Ahmed Abdel-Lateif. "Comparison of the Mortality Prediction of Different ICU Scoring Systems (APACHE II and III, SAPS II, and SOFA) in Acute Respiratory Distress Syndrome Patients." Chest 149.4 (2016).

[26] Sadaka, Farid, Margaret Cytron, Kimberly Fowler, Victoria Javaux, and Jacklyn O'Brien. "996: A COMPARISON OF APACHE II AND APACHE III SCORES IN PREDICTING MORTALITY OF PATIENTS WITH SEPSIS." Critical Care Medicine 42 (2014).

[27] Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS: A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit Care Med 1985, 13:519-525.

[28] Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) Based on an International Cohort of Intensive Care Unit Patients. JAMA. 1993;270(20):2478-2486. doi:10.1001/jama.1993.03510200084037.

[29] Rapoport J, Teres D, Lemeshow S, Gehlbach S: A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. Crit Care Med 1994, 22:1385-1391.
[30] Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA: Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). Crit Care Med 2007, 35:827-835.

[31] Nathanson BH, Higgins TL, Teres D, Copes WS, Kramer A, Stark M: A revised method to assess intensive care unit clinical performance and resource utilization. Crit Care Med 2007, 35:1853-1862.

[32] Lemeshow S1, Klar J, Teres D, Avrunin JS, Gehlbach SH, Rapoport J, Rué M: Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective, multicenter study. Crit Care Med. 1994 Sep;22(9):1351-8.

[33] Le Gall J-R, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R: A simplified acute physiology score for ICU patients. Crit Care Med 1984, 12:975-977.

[34] Ohno-Machado, L, Resnic, FS, and Matheny, ME (2006). Prognosis in critical care. Annual review of biomedical engineering, 8(1981):567-99.

[35] Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR: SAPS 3 - from evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med 2005, 31:1345-1355.

[36] Metnitz PG, Valentin A, Vesely H, Alberti C, Lang T, et al. 1999. Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Simplified Acute Physiology Score. Intensive Care Med. 25(2):192–97

[37] Vincent JL, Moreno R, Takala J, Willatts S, Mendonca A, Bruining H, Reinhart C K, Suter PM, and Thijs LG (1996). The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Medicine, 22(7):707-710.

[38] L.A. Zadeh, “Discussion: Probability theory and fuzzy logic are complementary rather than competitive”, Technometrics, 37(3), 1995.

[39] A.H. Meghdadi, M.R. Akbarzadeh-T, "Probabilistic fuzzy logic and probabilistic fuzzy systems", *The 10th IEEE International Conference on Fuzzy Systems*, vol. 3, pp. 1127-1130, 2001.

[40] Mendonca LF, Vieira SM, and Sousa JMC, “Decision tree search methods in fuzzy modeling and classification,” International Journal of Approximate Reasoning, vol. 44, no. 2, pp. 106–123, 2007.

[41] Waltman L, Kaymak U, and Van den Berg J. Maximum likelihood parameter estimation in probabilistic fuzzy classifiers. In Proceedings of the 14th Annual IEEE International Conference on Fuzzy Systems, pages 1098–1103, Reno, Nevada, USA, 2005.

[42] Van Den Berg Jan, Uzay Kaymak, and Rui Jorge Almeida. "Conditional Density Estimation Using Probabilistic Fuzzy Systems - IEEE Xplore Document." IEEE TRANSACTIONS ON FUZZY SYSTEMS. N.p., 21 Dec. 2012. Web. 01 Apr. 2017.

[43] Zadeh L. Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications*, 23(2):421–427, 1968.

[44] Van den Berg J, Kaymak U, and Van den Bergh WM, “Fuzzy classification using probability-based rule weighting,” in Proc. 11th IEEE Int. Conf. Fuzzy Syst., 2002, pp. 991–996.

[45] Van den Berg J, Kaymak U, and Van den Bergh WM. Financial markets analysis by using a probabilistic fuzzy modelling approach. *Int J Approx Reason*, 35:291–305, 2004.

[46] Karayiannis NB and Mi GW. Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques. IEEE Transactions on Neural Networks, 8:1492–1506, 1997.

[47] Van den Berg J, Van den Bergh WM, and Kaymak U, “Probabilistic and statistical fuzzy set foundations of competitive exception learning,” in Proc. 10th IEEE Int. Conf. Fuzzy Syst., 2001, pp. 1035–1038

[48] Navot A. On the Role of Feature Selection in Machine Learning Thesis. Senate of the Hebrew University. December 2006.

[49] Hughes G. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory, 14(1), pages 55-63. January 1968.

[50] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), pages 2507–2517. 2007.

[51] Guyon I, Gunn S, Nikravesh M, and Zadeh LA, editors. Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing). Springer. August 2006.

[52] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, and Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology, 21(1):128–138, 2010.

[53] Abdala OT, Saeed M. Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted K-nearest neighbors algorithm. Comput. Cardiol. 2004; 31:693–696.

[54] Clifford GD, Long WJ, Moody GB, and Szolovits P. Robust parameter extraction for decision support using multimodal intensive care data. Philosophical transactions, Series A, Mathematical, physical, and engineering sciences, 367:411–429, 2009.

[55] Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JMC, and Finkelstein SN (2013). Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine.*

[56] Paetz J. Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions. Artificial Intelligence in Medicine 2003;28(6):207–30.

[57] Cismondi F, Fialho AS, Vieira SM, Sousa JMC, Reti SR, Howell MD, et al. Computational intelligence methods for processing misaligned, unevenly sampled time series containing missing data. In: CIDM. 2011. p. 224–31.

[58] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS, Escobar GJ, Angus DC. Assessment of Clinical Criteria for Sepsis For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-774. doi:10.1001/jama.2016.0288

[59] Shankar-Hari M, Harrison DA, Rowan KM. Differences in Impact of Definitional Elements on Mortality Precludes International Comparisons of Sepsis Epidemiology-A Cohort Study Illustrating the Need for Standardized Reporting. Crit Care Med. 2016;44(12):2223-30.

[60] Takagi T and Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man and Cybernetics, 15(1):116–132, 1985.

[61] Slesinger Todd L and Laurence Dubensky. Sepsis – 3, a New Definition. Solutions or New Problems? American College of Emergency Physicians, July 2016.
[62] Steven Q. Simpson, New Sepsis Criteria, Chest, Volume 149, Issue 5, 2016, Pages 1117-1118, ISSN 0012-3692.

[63] Rymson M. Balentien. Personalized mortality risk predictions for ICU Septic Shock patients: Making use of Probabilistic Fuzzy Systems. Master's dissertation. July 2016.

[64] A.S. Fialho, S.M. Vieira, U. Kaymak, R.J. Almeida, F. Cismondi, S.R. Reti, S.N. Finkelstein, J.M. Sousa. Mortality prediction of septic shock patients using probabilistic fuzzy systems. Applied Soft Computing, 42 (2016), pp. 194-203.

[65] P. Somol, B. Baesens, P. Pudil, J. Vanthienen. Filter- versus wrapper-based feature selection for credit scoring International Journal of Intelligent Systems, 20 (2005), pp. 985-999.

[66] Y. Wang, J. Wang, H. Liao, H. Chen, An efficient semi-supervised representatives feature selection algorithm based on information theory, Pattern Recognition, Volume 61, 2017, Pages 511-523, ISSN 0031-3203.

[67] Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine, 4(2), 627–635.

[68] E. Hanisch, R. Brause, J. Paetz, and B. Arlt. Predicting death for patients with abdominal septic shock. Journal of Intensive Care Medicine, 26:27–33, 2009.

[69] Maastricht University Medical Center+. Healthy Living: Strategy for 2020. January 2015. Print.

[70] Timsit JF, Fosse JP, Troche G, De Lassence A, Alberti C, Garrouste-Orgeas M, et al. Accuracy of a composite score using daily SAPS II and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 h. Intensive Care Med. 2001;27(6):1012–21. doi: 10.1007/s001340100961.

[71] van Zanten, A. R., Brinkman, S., Arbous, M. S., Abu-Hanna, A., Levy, M. M., & de Keizer, N. F. (2014). Guideline bundles adherence and mortality in severe sepsis and septic shock. Critical care medicine, 42(8), 1890-1898.

[72] Arabi, Y., Al Shirawi, N., Memish, Z., Venkatesh, S., & Al-Shimemeri, A. (2003). Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study. Critical care, 7(5), R116.

[73] Minne, Lilian, Ameen Abu-Hanna, and Evert de Jonge. "Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review." Critical care 12.6 (2008): R161.

[74] Safari, Saeed, et al. "Accuracy of SOFA score in prediction of 30-day outcome of critically ill patients." Turkish journal of emergency medicine 16.4 (2016): 146-150.

[75] Moreno, Rui P., et al. "SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission." Intensive care medicine 31.10 (2005): 1345-1355.

[76] Warren, James, Gleb Beliakov, and Berend Van Der Zwaag. "Fuzzy logic in clinical practice decision support systems." System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on. IEEE, 2000.

[77] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

[78] Hand, David J., and Robert J. Till. "A simple generalisation of the area under the ROC curve for multiple class classification problems." Machine learning45.2 (2001): 171-186.

[79] Hamel, Lutz. "Model assessment with ROC curves." Encyclopedia of Data Warehousing and Mining, Second Edition. IGI Global, 2009. 1316-1323.