Risk assessment and mortality prediction in patients with venous thromboembolism using big data and machine learning

A thesis submitted in fulfilment
of the requirements for the degree
of Master by Research
at Bournemouth University
by

VASILIKI DANILATOU, MD, PhD

Bournemouth
University

December 2020

Acknowledgements

First of all, I would like to express my gratitude to my supervisors, Prof. Theodoros Kostoulas and Prof. Vasilios Katos for their continuous guidance and financial support for my studies. They helped me with patience in confronting difficulties and disappointments during the course of my studies, since my background studies were based on medicine and clinical and laboratory practice and I had to manage and understand a multidisciplinary project. This work was supported partially by IDEAL-CITIES; a European Union's Horizon 2020 research and innovation staff exchange program (RISE) under the Marie Skłodowska-Curie grant agreement No 778229.

I also feel grateful to all of those with whom I had the pleasure to collaborate during this period and exchange fruitful experiences and ideas. This work would not have been possible without the help of my esteemed collaborators from FORTH, Dr. Despoina Antonakaki, Dr. Alexandros Kanterakis, and Dr. Christos Tzagkarakis. They provided me significant help with coding and ML, areas that I found to be the most challenging aspects of my thesis in the outset. Through this collaboration and journey, I gained valuable computing skills that I can now confidently apply in the domain of medical sciences. I am ever so grateful that this knowledge will not only offer me new opportunities in my career, but will enable me to work on the forefront of medical research and contribute to my profession. I would also like to express my respect to my director Aikaterini Sfiridaki and my colleagues from General Hospital Venizeleio for their patience and understanding of my pressure, as I juggled between my studies, work in the hospital, on-call services and family for the past two years. In addition, at the beginning of the project I had the opportunity to share my thoughts and scientific plans with Prof. James Douketis from McMaster University, Canada as well as Prof. Grigoris Gerotziafas from University Pierre et Marie Curie, France, that were really helpful and supportive and provided me with useful information.

I would also like to thank Prof. Ioannis Tsamardinos for offering JADBIO and his team, Vincenzo Lagani and Naomi Thomson for their help and guidance in using this user-friendly web-based tool.

Last but not least, I would like to thank my husband Prof. Sotirios Ioannidis, as without his support, encouragement and enthusiasm of him I wouldn't dare to start such a challenging project and of course my beloved kids, Jason, Ariadne and Io for their patience and understanding during this hard-working period. Their love and support kept me motivated to reach my goals.

Abstract

Venous thromboembolism (VTE) is the third most common cardiovascular condition that affects mainly hospitalized and cancer patients and it is associated with high morbidity and mortality. Some patients need immediate treatment and monitoring in intensive care units (ICU). Moreover, cancer patients are at increased risk of developing VTE, especially in the immediate period after ICU hospitalization. It is crucial to predict which of the cancer patients will develop VTE, as well as early and late mortality in these high-risk patients and recognize possible treatable factors in order to improve survival. Several scoring and predictive models have been developed for these purposes, but with limited generalizability and they are mostly effective in the prediction of in-hospital mortality. They have several limitations, for example they use data recorded only on the first day of admission. Moreover, no score exists so far to predict late mortality in ICU patients. With the advanced use of electronic health records, open-source big-data medical databases and machine learning, predictive modelling could be utilized and become a powerful tool to guide clinical decision.

The aim of the study was to explore the use and performance of various machine learning algorithms (ML) in order to predict two research questions: (i) VTE risk in ICU hospitalized cancer patients after discharge and, (ii) early and late mortality in VTE patients hospitalized in ICU. For that reason, a freely accessible database MIMIC-III has been used that contains a vast amount of various time-series healthcare data from thousands of patients, making it ideal for ML based forecasting. Since it provides information even after discharge from ICU, it gives an opportunity to predict late mortality. Two groups of datasets were extracted from the database: D1, consisted of 4,699 patients with cancer who were admitted to ICU and stratified in two groups based on whether they were readmitted to ICU within 90 days with a diagnosis of VTE or not. The ML classification task was to predict which of the cancer patients originally admitted to ICU will be readmitted with VTE within 90 days. D2, consisted of 2,468 patients who were admitted to ICU with a VTE diagnosis and stratified in three groups, based on their outcome, that is, died during their first ICU admission (early mortality group), died after their discharge from ICU or in a later admission (late mortality group) and remained alive for months after their admission in ICU. In

this case, two ML classification tasks were constructed, first to build a model that distinguishes early mortality and second, a model that distinguishes late mortality.

A very wide range of features were selected, that includes demographic information, clinical and laboratory data, prescriptions, procedures, well established comorbidity and severity scores as well as information coming from written notes. Clinically relevant entities from free medical notes were extracted using the sequence annotator SABER and then they were fitted into a Latent Dirichlet Allocation (LDA) model of 50 topics. In total, 1,471 features were extracted for each patient, grouped in 8 categories, each representing a different type of medical assessment. Automated ML platform that easily handles with-high dimensional, noisy and missing data, as well as Monte Carlo simulations based on Random Forests with hyperparameter tuning and class-balancing with Synthetic Minority Oversampling Technique (SMOTE) were trained in parallel.

Due to the highly imbalanced nature of the first dataset (“cancer patients with thrombosis”), neither of the ML approaches were able to predict DVT in cancer patients even after the use of SMOTE method. As far as it concerns the prediction of early mortality in ICU patients with VTE, the best ML model chosen to predict early mortality was Random Forests (AUC=0,92). Regarding late mortality, the best ML model was again Random Forests. Nevertheless, the task of predicting late mortality was less efficient even with the holistic approach (AUC=0,82). Significant clinically relevant predictive features of early and late mortality were cancer, age, treatment with warfarin, and red cell transfusions, whereas known severity scores performed well only in the prediction of early mortality.

The contribution of this study to the current knowledge was multi-leveled, as it explored the performance of various ML approaches in a big-data driven research approach, using multiple formats of data from structured to unstructured medical notes, it examined the effect of balancing techniques in highly imbalanced datasets, such as the case of medical datasets, and finally discovered possibly new biomarkers. Early mortality in critically-ill patients with VTE can be easily predicted by ML techniques, whereas in the case of late mortality, which is a more difficult task, and where medical scores are still lacking, ML could probably outperform classic statistical methods. There is a need for more precise and reliable tools in order to overcome the nature of highly imbalanced medical datasets, such as the case of “cancer patients with thrombosis” dataset. This study showed that automated ML approaches have similar performance with manual selection and parametrization of ML models, which is highly promising in the setting of healthcare “big-data” medical databases.

TABLE OF CONTENTS

1 INTRODUCTION

Venous thromboembolism (VTE) is a potentially lethal disease that presents with clots in the veins, most frequently as deep vein thrombosis (DVT) and pulmonary embolism (PE). It is a quite common problem with an annual prevalence rate of approximately 1 per 1000 adults [1]. Its prevalence has been reported to increase probably due to a doubling of life expectancy and quadrupling of the world population during the 20th century[2]. The impact of this disease is enormous since it has severe physical and psychological complications, such as post-traumatic stress disorder, post-thrombotic syndrome, recurrence, and even death. More specifically, post-thrombotic syndrome impairs negatively the quality of life, and increases the healthcare costs[3]. Thrombo-embolic disease is one of the main causes of mortality in the world as it is estimated that it accounts for 1 in 4 deaths worldwide in 2010[4]. Its prevalence is even higher in hospitalized, critically-ill and cancer patients[5,6,7]. VTE in critically-ill patients is associated with significant morbidity, prolonged intensive care unit (ICU) and hospital stay and increased mortality[8]. For these reasons, it is crucial to predict promptly which patients are at high risk, as well as in-hospital and later mortality, and potentially identify new predisposing factors.

VTE is a complex multifactorial disease. Both acquired and hereditary factors interact and play essential roles in its development and outcome. The acquired risk factors can be transient or permanent depending on how long they persist. Based on their predictive value, they can be further stratified as strong (odds ratio >10), moderate (odds ratio 2–9), and weak (odds ratio <2). Examples of strong risk factors are orthopaedic surgery, major general surgery and major trauma. Moderate risk factors include central venous catheters, congestive heart or respiratory failure, cancer, chemotherapy, hormone replacement therapy, oral contraceptive therapy, and pregnancy/postpartum. Whereas bed rest (>3 days), air travel >8 hours, increasing age (≥40 years), and obesity are considered as weak risk factors [9,10]. Inherited factors are also classified as strong, medium and weak. Deficiencies of some natural coagulation inhibitors including antithrombin, protein C, and its cofactor protein S belong to strong genetic risk factors, as well as homozygosity of factor V Leiden (FVL) causing resistance to activated protein C, homozygosity of prothrombin G20210A which results in increased prothrombin levels and double heterozygosity of these mutations. Moderate genetic risk factors consist of heterozygous mutation in the FVL or prothrombin G20210A, and blood group (non-O blood group). Weak risk factors are considered hyper-homocysteinemia and homozygosity for factor XIII 34Val alleles [10,11]. The above-mentioned classification schema is not widely accepted and probably of low clinical importance since guidelines use different classifications, there are broad confidence intervals of risk estimates and the risk of thrombosis depends on more complex gene-gene and gene-environment interactions[12], but it could be a baseline approach in risk stratification.

VTE is also a frequent complication in patients with active cancer. Cancer itself increases directly and indirectly thromboembolic risk by various pathophysiological mechanisms. Cancer cells secrete inflammatory cytokines and micro-particles, directly activate coagulation mechanisms and platelets leading to a prothrombotic state. Moreover, hospitalizations, surgical interventions, chemotherapy, the presence of central venous catheters, as well as the type and stage of cancer, the presence of comorbidities and advanced age are important predisposing superimposed factors. It is crucial for clinicians to prevent thrombosis in these high-risk patients as well as to realize that prevention is a life-saving procedure, since VTE development during the first year from diagnosis of cancer increases mortality and affects negatively the outcome of disease [13].

VTE could be prevented if prompt and accurate selection of patients at high risk of thrombosis and prophylactic anticoagulation are applied. Unfortunately, there is no such a simple and straightforward method to predict thrombosis. Clinicians in their every-day clinical practice are constantly confronted with the dilemma of prophylactic anticoagulation in high-risk patients, since the balance of risks between thrombosis and bleeding cannot be quantified by clinical experience and most frequently there is a tendency to overestimate bleeding risks[14]. Moreover, recent negative personal experiences can affect objective judgment. To overcome this difficulty, several risk assessment models (RAMs), scores and tools such as Khorana [15] and COMPASS-CAT [16] score have been developed to predict thrombo-embolism in hospitalized or ambulatory cancer patients respectively, but they have so far limited generalizability and validation[15,16]. External validation in large data sets is always necessary before these tools can be broadly implemented

[17]. The risk stratification in cancer patients has been problematic due to the broad heterogeneity of different cancers, the uniqueness of different patients and the coexistence of various pathologies that predispose both to increased bleeding and thrombotic risk.

Some high-risk patients that present with thrombosis need immediate hospitalization in ICU and suffer from high mortality incidence. There are several scores to predict mostly in-hospital mortality and early mortality in ICU patients. The Simplified Acute Physiology Score (SAPS)[18], Acute Physiology and Chronic Health Evaluation (APACHE)[19] and Sequential Organ Failure Assessment (SOFA)[20] score, are based on patient measurements during the first 24 hours of hospitalization and are considered validated tools in predicting early mortality[21]. On the other hand, long-term survival after ICU admission is not well studied and risk assessment models are missing so far. It has been recognized that this is an important outcome that needs to be accurately predicted and prevented, since it could assist difficult clinical decision making and improve medical costs[22]. For example, more accurate estimates of long-term outcomes at the individual level, could assist clinicians in important decisions regarding rational allocation of the limited medical resources, an important consideration especially in the era of COVID-19 pandemic.

Nevertheless, traditional RAMs have several limitations. They have been developed based on different target populations with heterogeneous inclusion and exclusion criteria, thus during validation they provide modest performance. For example, the accuracy of various scores drops in the elderly population[23], since there is a significant correlation of various parameters with age (e.g. D-dimer and age correlation). Moreover, they are based on multivariate statistical methods, such as logistic regression models, that disregard the non-linear relationships that exist between variables in real medical datasets. These scores are built based on health data collected during the first 24 hours of ICU admission or instant based measurements (e.g. the worst or average value), and do not consider time-series measurements, that could contain important information for clinical deterioration[24]. Changes of organ function variables over time could provide more useful information with greater prognostic relevance. Simplified integer-based scoring systems neglect the complex nature of variables (for example hypertension could both increase thrombotic as well as bleeding risk). Moreover, it has been reported variable interobserver agreement in the application of these scores based on the personal experience of clinicians, so there is a possible bias in the interpretation of RAMs[25]. Another significant problem is the use of different laboratory methodologies with varying specificities, sensitivities and cut-off values that produces difficulties in the comparison between various facilities[26].

As an adjunct in the above-mentioned problems, there has been an increasing interest in the use of machine learning (ML) approaches in the prediction of various outcomes in medicine[27], since ML could recognize complex pattern changes in data and associations, that could probably help in improving patient care and survival, as well as lower hospitalization costs. On the other hand, the growing availability of large-scale healthcare big data and automated patient surveillance systems could improve clinical decision-making[28]. These data are not only large in size and dimensionality but also unstructured and heterogeneous. Using a holistic approach, incorporating large scale healthcare data could advance personalized and precision medicine.

This study focused on the exploration of automated (autoML) as well as custom ML algorithms in the prediction of two important clinical questions, such as mortality and thrombosis in ICU hospitalized patients. A holistic approach was used choosing a high dimensional dataset, with thousands features of various formats, and further processing has been applied to manage a high imbalance ratio with the final goal to improve performance of the proposed model. More importance has been given to the collection and combination of a very wide selection, but thrombosis-oriented of heterogeneous clinical and laboratory features as well as free-text medical notes. Data were identified and selected retrospectively over a period of time and hospitalized ICU patients had a long-term follow-up in the database. The initial hypothesis was that use of multiple ML algorithms could outperform existing prognostic scores, as well as refine them by identifying new biomarkers. Finally, an effort towards selecting important clinical features has resulted in clinically meaningful bio-signatures. This study using a novel approach that exceeds the classic statistical methods, has contributed in the prediction of early and late mortality in ICU-hospitalized patients with thrombosis, the identification of bio-signatures and rediscovery of candidate new biomarkers using “big-data”, combined with medical expertise and ML approaches.

2 AIM OF THE STUDY

Given that there is no universal consensus in the use of a specific predictive score in patients with cancer and/or thrombosis and that scores are not rigid and are highly subjective, this study aims to explore the usage, applicability and performance of machine learning algorithms in a big-data driven research approach, to answer two important research questions:

(i) Is it possible to use ML in prediction of VTE-associated readmission of ICU hospitalized cancer patients, after discharge?

(ii) Is it possible to predict early and late mortality in VTE patients hospitalized in ICU?

To fulfil these goals, the following objectives must be met:

1) Data acquisition and definition: To correctly assess VTE risk and predict outcome in ICU hospitalized patients it is necessary to have a wide range of high-quality and high-frequency medical data. Attributes must be carefully selected according to current knowledge to avoid noise and "garbage-in, garbage-out" effects. Multiple different formats of data need to be processed in a homogeneous pattern (e.g. conversion of textual information to numerical and extraction of meta-features).
2) Application of ML method and model training: Identification of best ML algorithms is time-consuming and needs extensive parametrization and grid-search. For these reasons, a dual approach will be used comparing automated with standard ML algorithms and hyperparameter tuning.
3) Implementation of balancing methods: Handling with highly-imbalanced data is a frequent problem in the medical field, thus impairing the performance of the proposed models. Exploration of balancing techniques could theoretically result in better performance.
4) Evaluation and interpretation of results: ML algorithms can be evaluated with standard statistical metrics. Besides that, an important challenge for medical researchers is that ML algorithms results, ideally must be explainable in order to identify complex biological relationships and provide new insights. This would allow the identification of clinically meaningful predictive features that contribute to the predictive model.
5) Comparison with other RAMs: Comparison of the proposed framework with known Risk Assessment Models or published data.

3 REVIEW OF LITERATURE

Since the main research questions are focused on thrombosis prediction in critically-ill cancer patients and prediction of early and late mortality in ICU patients with thrombosis, in the following section a review of the existing risk assessment models for these two important clinical problems will be reported. These scores have been developed based on classic statistical methods. The novelty of the approach in the current study, is that to address these research questions, machine learning in big data will be used. In the following section of the review, background of using ML algorithms and automated ML platforms will be shortly addressed, as well as the importance of big-data in healthcare. Big data is a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional methods, but they are ideal for machine learning algorithms since the latter need large data for training. Finally, studies based on prediction of thrombosis using machine learning algorithms, as well as their limitations are discussed shortly.

3.1 Risk assessment models for prediction of thrombosis in cancer patients

Khorana score was the first tool that was developed to predict thrombotic risk in chemotherapy naïve patients[15]. It is simple in use but it has several constraints. Using simple laboratory parameters before chemotherapy treatment, patients are divided in three risk groups (low, intermediate and high) with a large proportion of them falling in the intermediate risk category, making debatable its clinical applicability. Moreover, it has low sensitivity in certain tumor types and this tool can be used only at diagnosis and before initiation of chemotherapy. To improve its predictive performance several modifications have been proposed but with limited generalizability. VIENNA-CATS score[29] improved the discrimination ability through addition of two biomarkers, D-dimers and P-selection, although the latter is a sophisticated test. PROTECHT score tried to expand Khorana score through incorporating specific types of chemotherapeutic agents that increase the thrombotic risk[30]. The ONCOTEV score [31] showed an improved discrimination accuracy of Khorana score by adding ultrasound in the diagnostic panel but it is still under validation. Recently a promising risk assessment tool, COMPASS-CAT derived from a large prospective cohort and focused in ambulatory cancer patients, has been shown to have improved sensitivity and specificity but it also needs further validation [16]. A direct comparison of different RAMs for VTE prediction in a cohort of lung cancer patients showed that the COMPASS-CAT model had an 100% predictive accuracy[32].

3.2 Risk assessment models for prediction of mortality in thrombosis patients

Another important clinical issue is the prediction of mortality in ICU hospitalized patients with thrombosis. Several prognostic models that incorporate clinical and or laboratory findings have been derived to predict early mortality in patients with thrombosis, such as the Pulmonary Embolism Severity Index (PESI) and the simplified PESI for pulmonary embolism which are the most well-known[33, 34 , 35]. Moreover, there are several other scores, such as SAPS [18] , APACHE [19], SOFA [20] , OASIS [36], that estimate the severity of disease in ICU and that correlate positively mostly with early mortality but have varying accuracy depending on the population studied. These scores are based on data obtained during the first day of admission or the worst value, so they lack considerable information stemming during their hospital stay and post-discharge. Their performance is lost over time, since medical practices change significantly. Moreover, they are not widely customized in different patient groups, such as patients with thrombosis or cancer. It should be noted that ICU patients are at increased risk of post-discharge morbidity and mortality. So far, accurate identification of patients who will stay at risk even months later is lacking.

It is crucial to predict these high-risk patients since proper screening or adequate treatment could probably improve their survival [37]. Moreover, all the above-mentioned tools were developed in an era without electronic health records, big data storage, and machine learning. In the last decades, there is an increasing interest in the use of information technology and ML algorithms in order to improve forecasting and possibly guide clinicians [27].

3.3 Basic ML algorithms background

Artificial intelligence (AI) is a system that has the ability to correctly interpret, learn from external data, and use them to achieve specific goals and tasks through flexible adaptive mechanisms like the human brain. AI is a research area which also deals with the interpretation of two types of data:

i. Structured, such as patient characteristics (e.g. demographic), laboratory and imaging data. These features can be either binary, categorical or continuous.

ii. Unstructured, such as clinical notes in the medical file or publications in medical journals.

Structured data can be analyzed by ML algorithms while natural language processing (NLP) can be used to extract information from unstructured data [38].

ML uses a combination of mathematics, statistics and computer science in order to achieve AI through learning from the available data, and thus the machine can be trained using the data and based on algorithms, gives the ability to learn how to perform a specific task. ML algorithms learn from a vast amount of input data (various patient features such as age, gender, body mass index, diagnosis) and they produce complex mappings between them in order to create an output (e.g. outcome of thrombosis or mortality). If the output is known this algorithm is called **supervised ML**, while if the output is unknown it is called **unsupervised**. Supervised learning performs better in predictive models since it can build relationships between inputs (patient traits) and output (outcome) but unsupervised learning could possibly discover unknown relationships or clusters of features. The goal of any supervised ML algorithm is to best estimate the mapping function for the output variable given the input data. The mapping function is often called the target function because it is the function that a given supervised ML algorithm aims to approximate. Different ML algorithms make different hypotheses about the form of the target function, for that reason it is necessary to try several algorithms in order to find the best for each function.

There are two types of algorithms, parametric and non-parametric. Parametric models summarize data with a set of parameters of fixed size, make large assumptions about the mapping of the input to the output variables, are simpler and faster to train, and require less data but may not be as powerful. Examples of parametric algorithms are Logistic Regression and Linear Discriminant Analysis. Nonparametric methods make few or no assumptions about the target function and require a lot more data, are slower to train and have a higher model complexity but can result in more powerful models. Examples of non-parametric methods are Decision Trees, Naive Bayes and Support Vector Machines (SVM)[39].

ML algorithms in some cases fail in the prediction process. There are two types of prediction errors, bias and variance error. **Bias** is the simplification of the assumptions made by a model to make the target function easier to learn. Generally parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems. **Variance** is the amount that the estimate of the target function will change if different training data are used. Ideally, it should not change too much between different training datasets. The ultimate goal of any supervised ML algorithm is to achieve low bias and low variance. In turn, the algorithm should achieve good prediction performance. Parametric ML algorithms often have a high bias and a low variance and the opposite applies for nonparametric algorithms. **Trade-off** is the strain between the error introduced by the bias and the variance. A common problem in ML that results in poor performance of the algorithm is **overfitting**. Overfitting happens when a model learns perfectly from the training data but cannot generalize to new data, resulting in a poor performance of the algorithm. To avoid overfitting, two methods exist, one is k-fold cross validation and the other is the partitioning of the data set to train and test validation set.

The most basic and simple ML algorithm is **linear regression** which is based primarily on statistics. Linear regression is a statistical model that assumes a linear relationship between one or more input variables (x, independent variables) and a single output variable (y, dependent variable). Linear regression has been used for predicting output variables with continuous values (regression problems). For example, for n number of predictors (x_1,x_2, …, x_n) the following regression equation takes place: $y=\beta_0+\beta_1x_1+…+\beta_nx_n+\varepsilon$ (where ε is the random error and β is the regression coefficients). Linear regression calculates the estimators or predicted weights of the regression coefficients (b_0, b_1, …, b_n) that they define the estimated regression function $f(\mathbf{x}) = b_0 + b_1x_1 + \cdots + b_nx_n$. Ideally the estimated or predicted response, $f(\mathbf{x}_i)$, for each observation i = 1, …, n, should be as close as possible to the corresponding actual response y_i. The differences y_i -

$f(\mathbf{x}_\mathrm{i})$ for all observations i = 1, …, n, are called the residuals and it is represented as the vertical distance between the line and the data points. Regression is about determining the best predicted weights, that is the weights corresponding to the smallest residuals. Linear regression is a popular statistical tool that has also been applied in ML but it has some limitations. Linear regression models use linear combinations of variables but in biology it has been demonstrated that interactions between variables are more complex and nonlinear [40].

Logistic Regression[41] is a statistical method for analyzing a dataset in which there are one or more independent variables (risk factors) that estimate the probability of an outcome to occur or not (in this case thrombosis or mortality), that is a classification problem. Logistic Regression works with binary data, where either the event happens (1) or not (0). In contrast with linear regression, logistic regression does not use linear relationships but the natural logarithm function to find the relationship between the variables and uses test data to find the coefficients. The function can then predict the future results using these coefficients in the logistic equation. Logistic regression uses the concept of odds ratios to calculate the probability. This is defined as the ratio of the odds of an event "happening" to "not happening". This method is quite easy and fast but is not suitable for high dimensional data[40].

Naive Bayes method[42] is a supervised learning algorithm that is founded on Bayes' theorem[43]. This theorem is based on conditional probability or the likelihood that an A event will happen given that another B event has already happened, as expressed in the following equation.

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(B|A)P(A)}{P(B)}$$

This algorithm is simple, requires little data but it assumes that the features being evaluated are independent of each other, an assumption that does not happen in real life[40].

Linear Discriminant Analysis [44] is a dimensionality reduction method. It is based on Naive Bayes theorem and can be applied when the outcome of classification is categorical and has more than two classes. The model assumes a Gaussian distribution of the input variables. Removing outliers and standardization of data (so that they have a mean of 0 and a standard deviation of 1 is considered helpful[40]).

K-nearest neighbour (K-NN)[45] is a non-parametric ML algorithm. Non-parametric algorithms do not require a certain distribution of the underlying data. This is particularly helpful in practice where most of the real-world datasets do not follow mathematical theoretical assumptions. It has been applied in pattern recognition, and data mining. To determine which of the K instances in the training dataset are most similar to a new input, a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance. Euclidean distance is calculated as the square root of the sum of the squared differences between point a and point b across all input attributes i [40].

$$EuclideanDistance\ (a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i\ )^2}$$

K -NN works well with a small number of input variables.

Decision trees[46] split the data multiple times according to certain cut-off values in the features. After splitting, different subsets of the dataset are created, with each instance belonging to one subset. The final subsets are called leaf nodes and the intermediate subsets are called internal nodes or split nodes. To predict the outcome in each leaf node, the average outcome of the training data in this node is used. Trees can be used for classification and regression problems and have been applied for decision support of medical practitioners. One of the most important drawbacks of classical decision tree algorithms is poor processing of incomplete, noisy data[47].

Support vector machines (SVM)[48] are supervised ML algorithms suitable for both regression and classification problems. Data are pointed in a space with n-dimensions (according to the number of features), and the most suitable hyperplane (decision boundary) that differentiates between the two classes is estimated. SVM algorithms use a set of mathematical functions that are defined as the kernel. Examples of used kernels are linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. SVM is effective in high dimensional data, but less efficient in large noisy data as it takes considerable training time.

Principal Component Analysis (PCA)[49] is a dimensionality reduction method for large datasets. For that reason, all variables are initially standardized according to the following equation[50].

$$z = \frac{value - mean}{standard\ deviation}$$

To remove redundant information correlation between input variables is identified with a covariance matrix. By computing the eigenvectors and eigenvalues (linear algebra concepts) from the covariance matrix it is possible to extract principal components. Principal components are new variables that are constructed from the initial variables either by mixture or linear combination and that have as much condensed information as possible. This is quite advantageous in the real-life dataset with thousands of features that intercorrelate.

Ensemble[51] methods are meta-algorithms that combine several machine learning algorithms into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking). A commonly used class of ensemble algorithms is Random Forests (RF) where bootstrapping is performed. Each tree in the ensemble is built from a sample drawn with replacement from the training set. In addition, instead of using all the features, a random subset of features is selected, further randomizing the tree.

Artificial neural networks (ANN)[52] is a network of ML algorithms resembling the human brain learning function through neurons. ANN can detect patterns and non-linear interactions in large complex data. A weight is placed on individual input data (input neurons) and they are fed in intermediate connections (hidden layers), and the interactions between the neurons are determined by optimizing the algorithm on large bodies of “training” data. During this process, multiple iterations are performed in which the properties of the neurons or nodes are adjusted in turn, and changes that improve the predictive power of the output are retained for the next iteration. Once trained, the neural network is then applied to previously unseen testing data, to assess its performance[53].

Reinforcement learning[54] is another category of ML which is similar to the Marcovian decision process[55] and uses interactions with the environment of reward or punishment type to make decisions[56].

Deep learning[57] uses many hidden layers of ANNs that process various information and stimuli from the surrounding environment. They have an excellent performance in complex tasks and in high dimensional data, they can learn and make decisions on their own but they are complex and not easy to understand [58]. An example of deep learning is Convolutional Neural Networks, the architecture of which is shown in *Figure 1*.

Natural language processing (NLP)[59] is a scientific topic that allow machines to extract information from text or speech. Sentiment analysis is one popular NLP tool that classifies texts into different categories relative to a positive, negative or neutral sentiment. A free-text is broken into smaller keywords or tokens of text (e.g. individual words) that can be used as features in an ML analysis.

A schematic representation of most commonly used ML algorithms is shown in Figure 1 provided by Rashidi et al [56]. In this study, supervised ML algorithms such as Logistic Regression, Decision Trees, Random Forests and Support Vector Machines were employed in the final ML pipeline, and NLP methods were employed, to extract information from clinical notes.

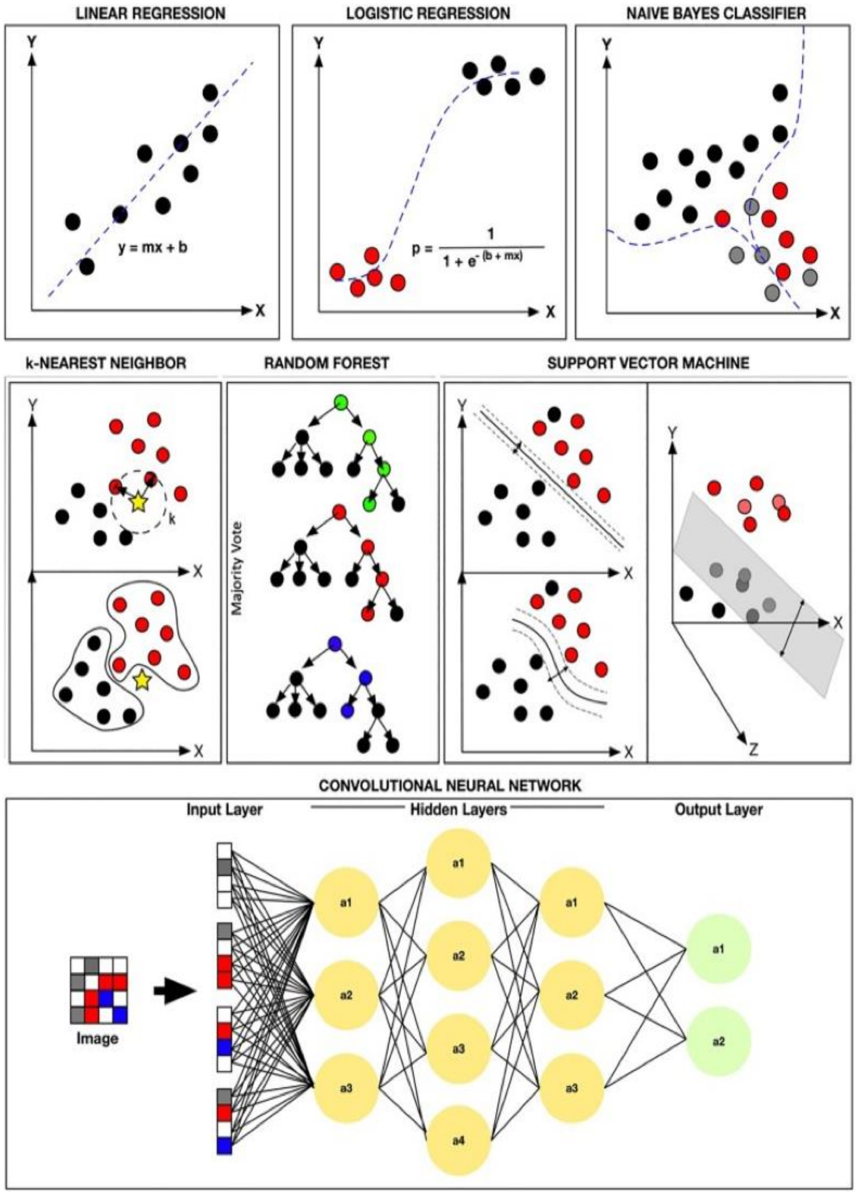

Figure 1. Schematic representation of the most commonly used ML algorithms (provided by Rashidi et al [56]).

3.4 Automated Machine Learning

The experimentation and extraction of the best performing ML model is time-consuming since it requires substantial human and computational effort, artificial intelligence expertise, and extensive tuning of hyperparameters. Moreover, the choice of algorithms and hyperparameter tuning is somewhat arbitrary, since they are difficult for humans to understand and they are treated as black boxes. For that reasons, several academia and industry based automatic ML tools have been developed to assist scientists (e.g. auto-WEKA [60], auto-sklearn [61]). An extensive comparison between auto ML platforms has been recently published[62]. AutoML is a rapidly developing field of ML. Moreover, the development of these autoML platforms provides a benchmark that will allow direct comparison and probably improved performance and reproducibility of the studies.

The basic pipeline of autoML approach has three steps: a) Data preprocessing and feature engineering, b) Model selection and hyperparameter optimization and c) Model interpretation and prediction analysis[63]. The first step is not yet developed fully in most autoML platforms, since considerable human interaction is needed in order to preprocess and transform data (e.g. conversion of categorical data into integers). After feature extraction is completed, the next step is training different types of models with hyperparameter optimization and selection of the best model (or an ensemble of models). Each platform uses a collection of known ML algorithms to build a model. For hyperparameter optimization, some of the most popular methods are grid search, random search, and Bayesian search. The third step, model interpretation is not supported yet from all autoML platforms.

3.5 Machine learning and risk assessment in the era of big-data

Most of the risk assessment models or prediction scores in medicine have been derived based on univariate and classic multivariate statistical analysis of collected data and selection of features that provide the best prediction accuracy methods[64]. Well established risk factors are included a priori but preliminary univariate analysis can reveal novel risk factors such as the case of platelet and leukocyte counts in Khorana score[15]. RAMs are originally structured to fit the derivation data set. Validation in independent test sets is always necessary but unfortunately these models do not perform as well during this second phase [17] .

Healthcare information has been overflowed by tons of data, such as electronic health records, freely accessible databases, genomic sequencing, medical imaging, wearable devices and smartphones, insurance and government records. The use of big-data analysis to deliver evidence-based information has been lagged so far, due to the difficulties in merging data into a common database and different types of format used. Several attempts so far to use big-data in healthcare involve data mining and analysis for diagnostic purposes, prevention of diseases, precision medicine, medical research, cost reduction and prediction of disease outcomes [65].

Given the fact that there is an increasing demand for “precision medicine” models, especially in oncology and with the growing availability of electronic health records (EHRs) and large healthcare databases (such as MIMIC III database[66], UK biobank[67]), new challenging opportunities are opened in medical research towards a “machine-learning” analysis of “big-data”. This approach exceeds the concept of classic statistical sampling and seems promising in risk assessment and prediction models.

Artificial intelligence and statistics differ substantially in their objective. ML models are designed for accurate predictions that can be generalized while statistical models are designed for inference about the relationships between different variables [68]. More specifically, inference corresponds to a mathematical model of the data generation process and formalizes the underlying system’s mechanism or tests a hypothesis about how the system behaves. Prediction aims at forecasting unseen data or future system’s behavior. Statistical models could be efficiently applied when the task at hand incorporates a tractable size (or dimension) of features and data size, while ML/AI could potentially fit better in problems with larger data size and high-dimensional feature space including non-linearities. To perform well, ML models generally need more data than statistical models. Limitations of statistical approaches (e.g. logistic regression) are, that they assume that features have a normal distribution and that a linear relationship exists between independent and dependent variables[69]. ML approaches have the advantage that they are not affected by bias and logic, they learn from big and complex data that a normal human brain cannot digest. The disadvantage of this process is that the machine cannot differentiate if an association reflects a true biological pathway [27]. In contrast to statistical methods, ML/AI methods usually have many hyper-parameters which need cautious tuning based on a training/test/validation/ dataset split, otherwise the performance of ML/AI model will be inferior.

ML/AI models could probably outperform RAMs by providing more accurate predictive results or possibly refine the parameters of medical scores. Only a few studies have recently tried to predict thrombosis using ML techniques, such as support vector machines or artificial neural networks [53 ,70 71]. Ferroni et al [70] used multiple kernel learning based on SVM and random optimization (RO) models to predict VTE risk in cancer patients. SVM is used to learn classifiers and RO to devise relative importance of different groups of clinical attributes in final predictions. The type of prediction is considered as binary since it is determined whether a patient will have a high risk of developing a VTE event in the future or not. VTE risk predictors are learned based on a 3-fold cross-validation on a training set that allows derivation of the model parameters. ML predictor outperformed Khorana score (AUC 0.716 vs 0.589). Qatawneh et al. [71] proposed a clinical decision system to automate and accurately predict the risk of VTE in hospitalized patients. They classified patients into five levels of risk based on predisposing factors chosen from the Caprini score of VTE model [72]. More specifically, the proposed approach is based on ANN in evaluating a multifactorial health issue. The system was developed a multilayered perceptron feed forward neural network which was trained using the Rprop training algorithm, and it consisted of an input layer with 35 neurons (representing the input variables for each patient such as age, gender, etc.), 3 hidden layers (where the number of neurons in the first, second and third hidden layer were 19, 10 and 5 respectively) and an output layer (that produced the type of the disease the patient suffered from). A stratified ten-fold cross validation was applied. This study was performed in only a few numbers of patients and appropriate metrics of performance are not reported.

Willan et al.[53] applied an ANN based method in order to risk stratify 11,490 patients referred with suspected DVT. This method could be extended for VTE prediction since it corresponds to a similar ML problem. More specifically, the authors introduced a system based on a standard binary classification problem, namely, whether or not the patient had a DVT. To address this, a standard binary-classification feed-forward artificial neural network was employed. The network consists of an input layer of 13 dimensions [sex, age, D-dimer result and the ten individual components of the Wells' score [73]], a hidden layer consisting of 8 neurons, and an output layer with one neuron. Each neuron contains a series of weights and biases which are multiplied and added to the inputs and then passed through an activation function that determines what numerical value is passed from a given neuron to the next layer or output from the network. It is these weights and biases that are optimized to obtain the best performance from the network in terms of DVT prediction. This study was designed as a proof of principle and the authors suggest that ANN could outperform existing scores of risk assessment such as Wells score, but they do not report metrics of performance and they concluded paradoxical clinical associations (e.g. they did not find an association of thrombosis with cancer or older age). Overall, prediction of venous thromboembolism with machine learning is limited so far, and current studies are sparse and problematic so further work is needed in that direction exploiting the advantage of big-data. *Table 1* summarizes the main characteristics and results of the above-mentioned studies. The only study that refers to cancer patients is by Ferroni et al.

Table 1. Studies that use ML algorithms to predict thrombosis

Authors	**Population studied**	**Attributes set**	**ML algorithm**	**Train/Test/ Validation**	**Perfor-mance metrics**	**Comparison with classic scores**
Ferroni et al [70]	1,179 ambulatory cancer patients	13	Multiple kernel ML (SVM and RO)	70/30	AUC: 0,716	Khorana (AUC:0,589)
Qatanweh et al [71]	150 hospitalized patient records	35 (based on Caprini score)	ANN (Multilayer Perceptron)	80/10/10	Recall: 80,7% Precision: 81,2%	Caprini score, no direct comparison
Willan et al [53]	7,080 eligible patients with suspected DVT	13 (including Wells score)	ANN	75/25	AUC: 0,89	Wells score included in the attribute set, no comparison

Abbreviations: ANN=Artificial Neural Network, AUC=Area under the curve, DVT= Deep Vein Thrombosis, RO= Random Optimization, ML= Machine Learning SVM= Support Vector Machine.

4 METHODOLOGY

4.1 Data source

Data were obtained from Medical Information Mart for Intensive Care (MIMIC-III, version 1.4) that is a large, freely-available database comprising of de-identified health-related data from 38,597 adult patients and 49,785 admissions in ICU of the Beth Israel Deaconess Medical Center, between 2001 and 2012. This database includes complex information such as demographics, time series measurements of vital signs (~1 data point per hour), laboratory tests, procedures, medications, caregiver notes, and mortality (including post-hospital discharge), as shown in *Figure 2* [66]. Clinical Classification Software (CCS)[74] is used to categorize diagnoses according to the International Classification of Diseases 9th edition (ICD-9 codes). Diagnosis is given as primary and secondary diagnosis ICD-9 codes as well as diagnosis-related groups (DRG)[75].

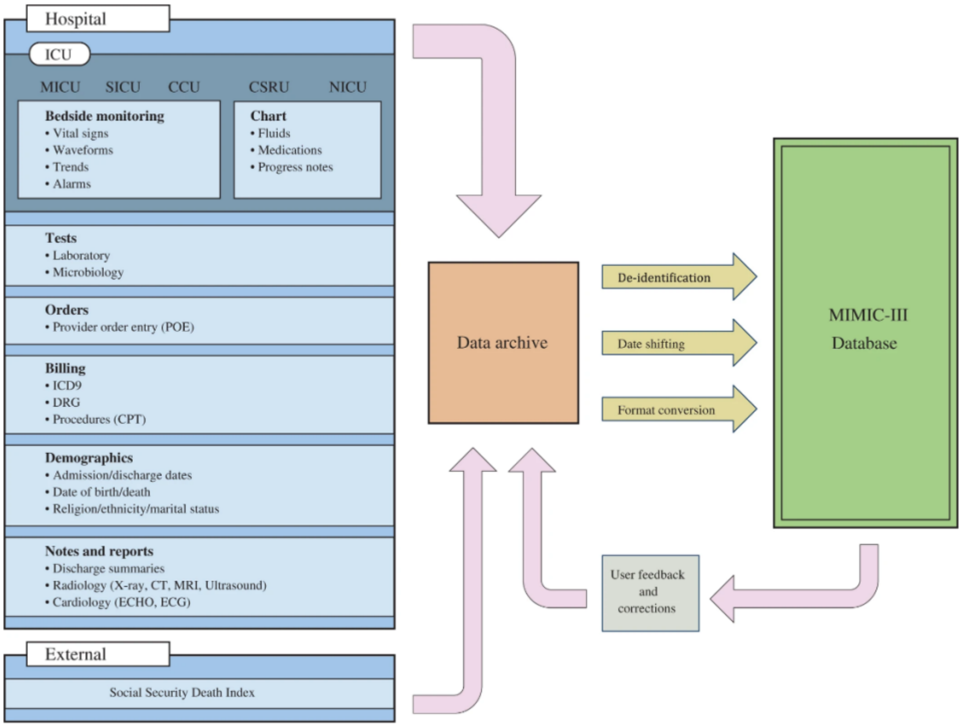

Figure 2. Overview of the MIMIC III database (provided by Johnson et al [66]).

Abbreviations: CCU=Coronary Care Unit; CSRU=Cardiac Surgery Recovery Unit; MICU=Medical Intensive Care Unit; SICU =Surgical Intensive Care Unit; TSICU= Trauma Surgical Intensive Care Unit.

4.2 Ethics statement

The MIMIC-III database was created in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards and data access was approved by PhysioNet (account credentialised on September 17, 2019). Patient data are de-identified and date-shifted. All pre-processing and data analysis were performed under MIMIC-III regulations.

4.3 Dataset description

Two datasets D1 and D2 were extracted in order to develop models for the two prediction tasks. D1 is identified as the dataset of patients with cancer that subsequently readmitted to ICU with a primary diagnosis of VTE within 90 days from the first ICU admission. This dataset was used in order to predict VTE risk in ICU hospitalised cancer patients after discharge. D2 is the dataset of patients admitted in ICU with a primary diagnosis of VTE. This dataset was used for predicting early and late mortality in VTE patients hospitalised in ICU.

For D1, 630 ICD9 codes were selected, related to common solid tumors and hematological malignancies that have increased thrombotic risk, i.e.gastrointestinal, urogenital, brain, breast, leukemias and lymphomas. For D1 and D2, 35 ICD9 codes related to deep vein thrombosis, thrombophlebitis and pulmonary embolism, were selected. Validation of this grouping for thrombosis diagnosis from an independent panel of physicians showed very good performance [76].

D1 database patient inclusion criteria: All patients aged >15 years old hospitalized in ICU with a primary diagnosis of cancer. Exclusion criteria: Age<15 years old (n=0), pregnancy and puerperium complications (n= 15), patients that presented with thrombosis in the first admission (n=527), patients with previous admission in ICU with thrombosis (n=36) patients with a subsequent thrombosis-related admission of more than 90 days (n=56), and patients with "do not resuscitate code" (DNR) (n=358). In total 5,691 cancer patients were identified (14,74% of total MIMIC-III patients). From this group of patients 4,642 did not develop thrombosis whereas only 57 cases of secondary thrombosis have been recognised with a median time to event of 36 days, mean 36,86 days (min 4 -max 85 days). The clinical characteristics of the D1 database are presented on *Table 2*.

D2 database patient inclusion criteria: All patients aged >15 years old hospitalized in ICU with a primary diagnosis of thrombosis. Three main diagnosis groups were identified, as shown in *Figure 3*: pulmonary embolism (n=960), deep vein thrombosis and thrombophlebitis (n=1543) and unusual site thrombosis (n=307). Many patients belonged in more than one diagnostic category. Exclusion criteria: Age< 15 years (n=3), pregnancy and puerperium complications (n=40) and patients with DNR (n=169). Overall 2,468 patients were selected (6.4% of total patients in MIMIC III) and split in 3 groups. The first, referred as G1 are 348 patients that died during the first ICU admission in which they were diagnosed with thrombosis. The second, referred as G2 are 817 patients that died after their discharge from ICU or in a later admission. On average this group died 549 days after admission with a median of 225 days. The third, referred as G3 are 1,303 patients that remained alive for months after their admission in ICU. From these groups two ML tasks were formed, the first is to build a model that distinguishes G1 vs. G3 patients (called “early mortality” or M1) and the second is a model that distinguishes G2 vs. G3 patients (called “late mortality” or M2). The clinical characteristics of the D2 database are presented on *Table 3***.**

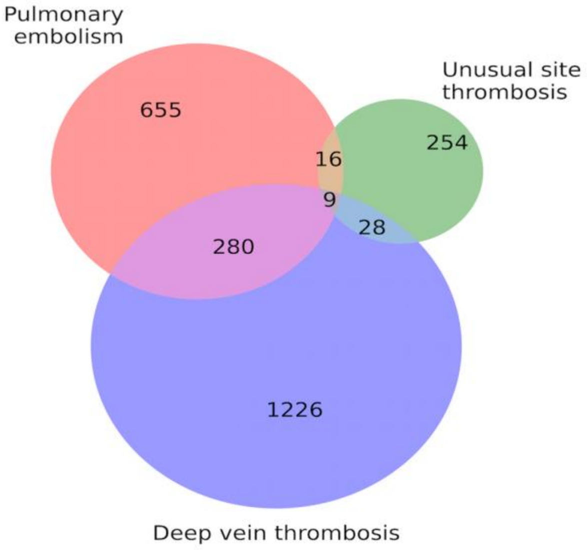

Figure 3. Venn diagram showing included and excluded cases from MIMIC III database in D2 database.

Table 2. Demographic and clinical characteristics of D1 database.

Characteristic	Number
Overall patients with cancer admitted in Intensive Care Units	5,691
- 1st admission with cancer never thrombosis	4,642 (81.5%)
- Readmission with thrombosis within 90 days	57 (1%)
Sex	
- Female	2,345(41.21%)
- Male	3,346 (58.79%)
Ethnicity	
- White	4,315 (75.82%)
- African/Americans	418 (7.34%)
- Other	958 (16.84%)
Age (years)	
- Average (Median)	66,19(66.89)
- min-max	18.87-98.86
Length of stay in days	
- Average (median)	16,68 (10,46)
- min-max	0-211,99
Number of admissions	
- Average (Median)	1,43(1)
- Min-max	1-10

Table 3. Demographic and clinical characteristics of D2 database.

Characteristic	Value	Characteristic	Value
Overall patients with thrombosis:	2,468	LOS, days	
- PE	960 (38.9%)	Average (SD):	7.06 (10.06),
- DVT	1,543 (62.5%)	Max length stay:	153.9 days
- Unusual site thrombosis	307 (12.4%)		
Sex		Number of admissions	
- Female	1,024 (41.5%)	- Average (SD):	1.15 (0.46)
- Male	1,444 (58.5%)	- Median:	1
Ethnicity			
- White	1,801 (73%)	Cancer diagnosis:	605 (24.5%)
- Black	246 (10%)		
- Other	421 (17%)		
Age, years		Mortality (%)	
Average (SD)	62,64 (16.7) [min=17.4 max=98.7]	G1 or Early (at the first admission):	348 (14.1%)
		G2 or Late (1-year mortality):	817 (33.1%)
		G3 or "Alive":	1,303 (52.8%)
		Time to death (in days)	
		Average (SD):	390 (647)
		Median:	83

Abbreviations: ICU=intensive care unit, LOS=length of stay, PE=pulmonary embolism, DVT=deep vein thrombosis, SD=standard deviation.

4.4 Attributes selection

For each of these patients a very wide selection of attributes (features) was extracted, selected manually based on factors that could be associated with thrombosis. In order to potentially investigate novel discriminatory attributes, a liberal approach on attribute extraction from the database was chosen, that is collecting as much as relevant data as possible. Data extracted included demographics (age, ethnicity), length of stay in ICU (in days), number of admissions, body weight, vital signs, basic laboratory indices (hematocrit, hemoglobin, white blood cells, platelets, renal and liver function tests, hemostasis screening tests, sepsis indices), severity scores, transfusion requirements, procedures, medications and mortality.

These attributes are grouped in 7 categories each representing a different type of medical assessment or intervention and one that included all features. The values of five of these were directly extracted from the corresponding tables of the database. These were LabEvents, that includes laboratory measurements, ChartEvents that includes charted data such as vital signs and blood pressure, InputEvents that includes transfusions and parenteral nutrition, Procedures and Prescriptions (medications). LabEvents were extracted in two values, the value of the first day and the average value (avg) during the admission. There are two types of InputEvents files MV, and CV since two different clinical information systems have been used, CareVue (Philips) and Metavision (iMDSoft). For these features the number of events and the overall received amount were recorded. 91 medications were extracted from Prescriptions and grouped in the following groups: vasopressors, antihypertensive, cardiovascular, antidiabetics, chemotherapy, growth factors, anticoagulants and antiplatelets.

4.4.1 Concepts

Concepts are meta-features containing the values of various scores. These values are not stored in the database but are available as SQL queries that estimate them from other features [77]. Concepts include a set of severity illness scores and organ failure scores such as Simplified Acute Physiology Score (SAPS), Sequential Organ Failure Assessment (SOFA), Glasgow Coma Scale (GCS), sepsis scores (Martin, Angus), first day laboratories, first day vital signs and transfusions.

It also includes comorbidities scores that are described as different Elixhauser indices [78]. Overall, 493 concepts were extracted.

4.4.2 NoteEvents

NoteEvents contain unstructured notes written by clinicians in free text format. Since one of the objectives of the study was to convert this textual information in numerical that could be added in the feature set, all clinically relevant entities from the text were extracted using the SABER sequence annotator [79] which is a Deep Neural Network framework, tailored for entity extraction from biomedical documents. SABER uses a Bi-directional Long Short-Term Memory (LSTM) architecture [80] [81] and provides access to pre-trained models for various types of entities. One of these is the disease ontology [82] [83](DO) which is a structured vocabulary of entities related to various pathologies and symptoms.

For each NoteEvent entry all DO entities were extracted, a process that required 30hrs in a computer equipped with 3 Nvidia GPUs, each with 16GB of memory. On average, for each patient 161 entities with a median of 133 were extracted. Next, these entities were fitted into Latent Dirichlet Allocation (LDA) topic model with the Gensim framework[84] by using 50 topics. LDA is a topic model that generates topics based on word frequency from a set of texts. A topic simply contains a probability distribution of entities, i.e. entity "pain", may belong by 20% in topic 1 and by 80% in topic 2. Ideally each topic is a thematic cluster that should contain entities with close semantic proximities, e.g. cardiovascular conditions (see *Figure 4*). Overall, this produced a 50-dimensional space that contained the topic distribution for each patient, or else, for each patient a vector of size 50 with thousands of topic marginal probabilities was obtained. For each patient, the extracted Disease Ontology tokens were projected into the 50-dimension Topic-Model space and this was used as NoteEvents features.

An example of the visualization of this model with the LDAvis tool[85] is shown in *Figure 4*. Principal Component Analysis on two dimensions was performed only for visualization purposes and this does not take any part in the text processing pipeline. The size of each topic (the circles) is relative to the sum of the absolute counts of the tokens that they contain. Overall, this process transformed the textual content for each patient in an easy-to-use numerical format that contained the basic thematic topics of these entries.

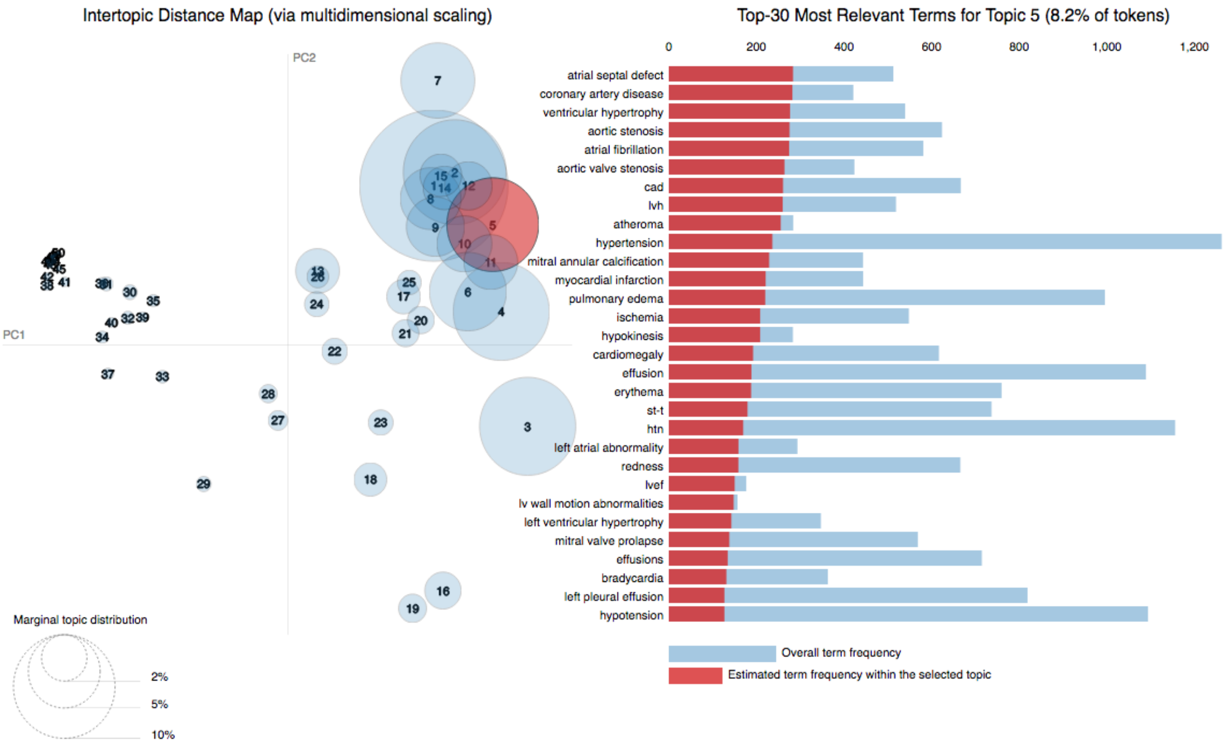

Figure 4. A visualization of the distribution of topics generated through the LDA topic modelling. Each circle on the left is a topic. The red circle is a random topic and the words on the right shows the relative distribution of its contained entities. In this example, the topic contains entities akin to cardiovascular conditions.

The overall number of features, the average and median number per patient, the most commonly found features in the patient group are described in details in *Table 4*. It is obvious that each group describes a different view of the clinical picture of the patient. Since one of the objectives of the study was to locate subsets of discriminatory features, a stratified analysis for each group was applied. Namely for each ML task, subsets were created that contained only the features of this group. Yet, all these subsets contained basic demographic information that are known to have strong correlation with mortality in thrombosis such as sex, length of stay and diagnosis group. Finally, a dataset that contained the entirety of the features was created. In total, 16 datasets were created, which correspond to the 2 ML tasks combined with the 8 groupings (7 groups plus 1 containing all groups).

Table 4. Description of clinical and laboratory features selected from MIMIC- III database. The first column described the corresponding table from the MIMIC-III database.

Group	Description	Features	Avg	Median	Most common features
Chart Events	Vital signs, labs, clinical information	235	433	77	Common labs, blood gases, blood pressure
Lab Events	Laboratory indices	45	1,237	1,157	Hematocrit, hemoglobin, white blood cells, platelets, red blood cells, renal and liver function tests, hemostasis screening tests, sepsis indices
Procedures	Several procedures including transfusion and mechanical ventilation	526	24.3	6	Venous catheterization, enteral nutrition, endotracheal intubation, mechanical ventilation for more than 96 hours
Input Events	Transfusion and parenteral nutrition	12 (MV) 10 (CV)			RBC transfusion, PLT transfusions, plasma transfusions
Prescriptions	Medications	91	132	14	Heparin, insulin, warfarin, aspirin, enoxaparin, norepinephrine, phytonadione and atorvastatin.
Note Events	Unstructured medical notes	50	48 entries, 2,408 characters	1,382 characters	N/A
Concepts	Scores, first day labs, first day vitals, doses and durations of medications	493			Comorbidity indices, severity illness scores, organ failure scores, sepsis scores, GCS, first day laboratories, first day vital signs, transfusions

Abbreviations: Avg=average, RBC=red blood cell, PLT=platelet, MV= Metavision, CV=CareVue, GCS=Glascow Coma Scale

4.5 Preprocessing

MIMIC III has applied an adjustment of the age in patients older than 89 years old to a fixed age of 300 years old, in order to adjust with privacy regulations. For that reason, these older patients were all assigned as 90 years old, given that risk of thrombosis is homogeneously high in ages more than 85 years old [86]. The Boolean values were replaced as TRUE:1, FALSE:0, and the gender (male/female) as well as the ethnicity (white/black/other) feature were one-hot encoded. Missing values are handled in two different ways. In the autoML approach preprocessing is automatically applied, by mean imputation and mode imputation, whereas in the custom approach a median imputation mode is adopted to fill the missing values.

4.6 Automated ML framework description

The AutoML platform, JADBIO uses an Artificial Intelligence (AI) Decision Support System called Algorithm and Hyper-Parameter Space selection (AHPS) in order to extract predictive models and signatures. It employs a recently developed protocol, namely Bootstrap Bias Corrected Cross-Validation (BBC-CV), for tuning the hyper-parameters of algorithms while estimating performance and adjusting for multiple tries. Standard preprocessing applied by JADBIO includes mean imputation, mode imputation, constant removal and standardization. JADBIO initially constructs a set of ML configurations consisting of algorithms and hyperparameters. The algorithms are Linear, Ridge and Lasso Regression, Decision Trees, Random Forests (RF) and Support Vector Machines (SVMs) with gaussian and polynomial kernels. This selection is based on the fact that these algorithms are most often the top classifier in extensive evaluation studies[87]. Subsequently it evaluates these configurations through bootstrap corrected cross-validation algorithm[88]. After selecting the "winning configuration" that is the best performing combination of preprocessing steps, feature selection algorithm and predictive algorithm that were tested during the analysis, it reports the classification statistics like truth table, AUC, sensitivity, specificity, precision, selected features along with their classification ability, sample predicted/real values. JADBIO applies all good practices of ML in order to eliminate any overfitting of the model and any bias in efficiency estimation. Details regarding the ML pipeline and statistical analysis can be found on [88]. Extensive testing showed that JADBIO's estimations lie towards the lower bound of the efficiency spectrum, or else these metrics are in fact conservative compared to the real classification ability of the generated model[89]. The user can select between three different types of analysis preliminary, typical and extensive with the latter extensively searching for an optimal model using high computational power. Another important and clinically relevant task of JADBIO is that it can identify biosignatures, that is a set of features with predictive ability, that could probably enforce knowledge discovery and further identify potentially new biomarkers.

4.7 Class imbalance

JADBIO addresses imbalanced classes through stratified cross-validation and diversified class weights during SVM learning. For that reason, it is crucial to examine the class balancing effect in light of oversampling combined with a state-of-the-art ML classifier, in this case RF classifier, which is robust and efficient when dealing with numerical, categorical and Boolean data. Towards achieving a balanced ratio between the two classes in both datasets, SMOTE method was adopted[90]. In particular, SMOTE generates synthetic minority class samples along the line segments joining randomly chosen m minority samples (i.e., m is the number of minority samples to oversample in order to obtain the desired class balancing ratio) and their K-nearest minority class neighbors. After defining m and K, SMOTE generates a new synthetic sample s of the form s=x+ρ(x-y), where x is the minority sample to oversample, y is one of its chosen nearest neighbors and ρ is a random number in the range of [0,1]. An increased generalization capability is expected, and thus an enhanced performance, of the used classifier since the generation of similar samples to the existing minority samples, creates larger and less specific decision boundaries. The default SMOTE implementation included in the Imbalanced - Learn Python package was used [91].

A shuffled stratified 75% train / 25% test split is applied on both datasets to divide it into a training and a test partition. Then, the training partition is divided into five stratified cross-validation folds (using shuffling). Since one of the objectives of the study was to examine SMOTE oversampling effect on the final performance evaluation, SMOTE was applied on all the “training” folds during each cross-validation iteration. The motivation towards applying oversampling during cross-validation is that similar patterns/instances may appear in both training and test partitions when the oversampling is performed prior to cross-validation which can lead to overoptimistic error estimates. However, if the oversampling is performed during cross-validation, only the training patterns/instances are considered both for generating new patterns/instances and training the model, alleviating over-optimism. In all cases, grid-search hyper-parameter tuning was performed as: the number of estimators [92] was selected out of this set: [10, 25, 50, 100], the maximum number of features was set as 'auto','sqrt' or 'log2', the maximum depth was selected from the set [10, 20, 30, 40], the minimum samples split [93] was selected from the set [5, 10, 15, 20] and the minimum samples leaf from[93] [2, 5, 10, 15].

The best hyper-parameters combination is computed according to an F1-score rule, i.e., the model selection is based on the highest F1-score on the “validation” fold for a specific hyperparameters combination. Then, the best (F1-based selected) RF model is trained on the entire initial (before the cross-validation iterations) training partition. Towards the final performance evaluation, the average ROC curves are computed, where the results are averaged over ten Monte Carlo repetitions with different realizations of the train/test split, the 5-fold stratified cross validation, and randomizations of the SMOTE method.

4.8 ML algorithm performance assessment

Performance of ML classification algorithms is typically assessed by simple statistical methods. Assessment of performance is done by the percentage of true predicted cases from the total cases. Sensitivity (or recall) is the proportion of true positives (true positives/actual positives or else $\frac{true\ positive}{true\ positive+false\ negative}$) and specificity the proportion of true negatives/actual negatives or else $\frac{true\ negative}{true\ negative+false\ positive}$ that are correctly identified. Accuracy is the proportion of the times which the classifier is correct, according to the following equation: $\frac{true\ positives+true\ negatives}{total\ predictions}$. Balanced accuracy is a better metric for imbalanced datasets, since it takes into account both positive and negative outcomes, according to the following equation:

$$(\frac{true\ positive}{true\ positive+false\ negative}+\frac{true\ negative}{true\ negative+false\ positive})/2 \text{ , or else } \frac{sensitivity+specificity}{2}$$

Precision is defined as the percentage of positive predictive values for each subject category. F1 score is the harmonic mean of the precision and recall, thus is another measure of test accuracy. Data are also represented in a confusion matrix as shown in *Table 5* . Receiver operating curves (ROC) illustrate the relationship between sensitivity (plotted on the y-axis) and specificity (x-axis). ROC curves can be easily interpreted by using area under the curve (AUC). AUC corresponds to the probability that a random sample would be correctly classified by each algorithm.

Table 5. A confusion matrix describes the performance of a classifier.

		Actual outcome	
		Negative	Positive
Predicted outcome	Negative	True negative	False negative
	Positive	False positive	True positive

5 RESULTS

5.1 Prediction of ICU readmission of cancer patients within 90 days due to thrombosis

Total number of patients included in dataset D1 is 4,699, where 4,642 patients have cancer and no thrombosis (cancer_never thrombosis) while 57 patients have cancer when admitted in ICU and then they develop thrombosis (cancer_then thrombosis), i.e., the imbalance ratio for D1 is 1: 81.28. As a result, D1 appears to be an extremely imbalanced dataset which is the basic reason for failing to achieve even modest mortality prediction results, using SMOTE and without explicitly adopting SMOTE technique. Similarly, JADBIO failed to predict accurately this event as shown in *Table 6*. Between 29,190 trained models the winning algorithm was Classification Random Forests training 1,000 trees with Deviance splitting criterion and minimum leaf size = 4. Among the most important features selected were concepts such as SOFA and sepsis Martin score, insertion of endotracheal tube, MCH (mean corpuscular hemoglobin, an index of red blood cells) and red blood cell transfusions (selected by Statistically Equivalent Signature algorithm with hyperparameters maxK=2, alpha=0.05).

In an effort to reduce the imbalance ratio and the dimensionality of the dataset, patients with cancer were narrowed down according to the ICD9 codes found in the thrombosis group. So, it was possible to reduce the size of the negative group (cancer_never thrombosis) to 2,937 vs 57 (cancer_then thrombosis), i.e. the imbalance ratio in this case id 1:51.5, which is slightly better but still high. Besides that, it was expected that the reduced dataset would be more homogeneous regarding cancer diagnosis. Moreover, features such as procedures with many missing values and NoteEvents were discarded, since they cannot be interpreted clinically, ending with 1,122 features (instead of the initial 1,471 features). Even with this modification it was impossible to improve performance of predictive algorithms (*Table 6*). As shown, Area under the ROC curve (AUC)[94] was 59%, which means the probability that the model ranks a random positive example more highly than a random negative example is 59%. The accuracy[95] or the fraction of the number of correct predictions to the total number of predictions, or else the fraction of predictions the model got right, has significantly improved from 11% to 98% in the reduced dataset, that is focusing in a more homogeneous patient group according to their diagnosis seems helpful. Also, F1 score[96] that measures the model's accuracy on the dataset, and is a combination of the precision and recall of the model, is quite low 0.04. Even if these results are not significant, it is quite interesting that the most important selected features in this case were again packed red blood cell transfusions, insertion of endotracheal tube and MCH.

Table 6. Detailed metrics of the performance for prediction of ICU readmission with thrombosis in cancer patients (within 90 days) using all features.

	ICU readmission with thrombosis in cancer patients (within 90 days)	ICU readmission with thrombosis in cancer patients reduced dataset
AUC [95% Confidence interval)	0.59 [0.50-0.69]	0.59 [0.46-0.7]
Accuracy	0.11	0.98
Balanced accuracy	0.52	0.5
F1 score	0.04	-
Precision	0.01	-
Sensitivity	0.93	-
Specificity	0.10	1
Average F1 score	0.52	-

5.2 Prediction of early and late mortality in ICU patients with thrombosis

5.2.1 Correlation of sepsis, comorbidities, and organ failure scores

Since MIMIC-III contains a variety of medical scores, the complex interactions between the parameters of various sepsis (n=20), comorbidities (n=17) and organ failure (n=12) scores were analyzed. For each of these score groups a Pearson pairwise correlation matrix was computed and these correlations were visualised with heatmaps. For sepsis and severity scores the "time before death" was added as an extra feature, which contains the negative of the time (in days) in which patients died after their first admission with a thrombosis diagnosis. For patients that were alive, this was left blank. Sepsis scores show a strong correlation between each other, as depicted in *Figure 5*. Surprisingly white blood cells, blood components transfusion and time before death do not seem to correlate well with sepsis. As far as it concerns comorbidity indices, presented in *Figure 6*, the Quan Elixhauer score was used, since both variants of Elixhauer measures AHRQ and Quan have comparable efficiency in predicting all-cause mortality [97]. Correlation between various diseases is shown, such as diabetes and renal failure, hypertension and renal failure, liver failure and alcohol abuse, congestive heart failure and chronic pulmonary disease, peripheral vascular disease and diabetes. *Figure 7* represents correlation between most important ICU severity scores. A strong correlation is observed between various severity and organ failure scores, although none of these scores showed a strong correlation with time before death.

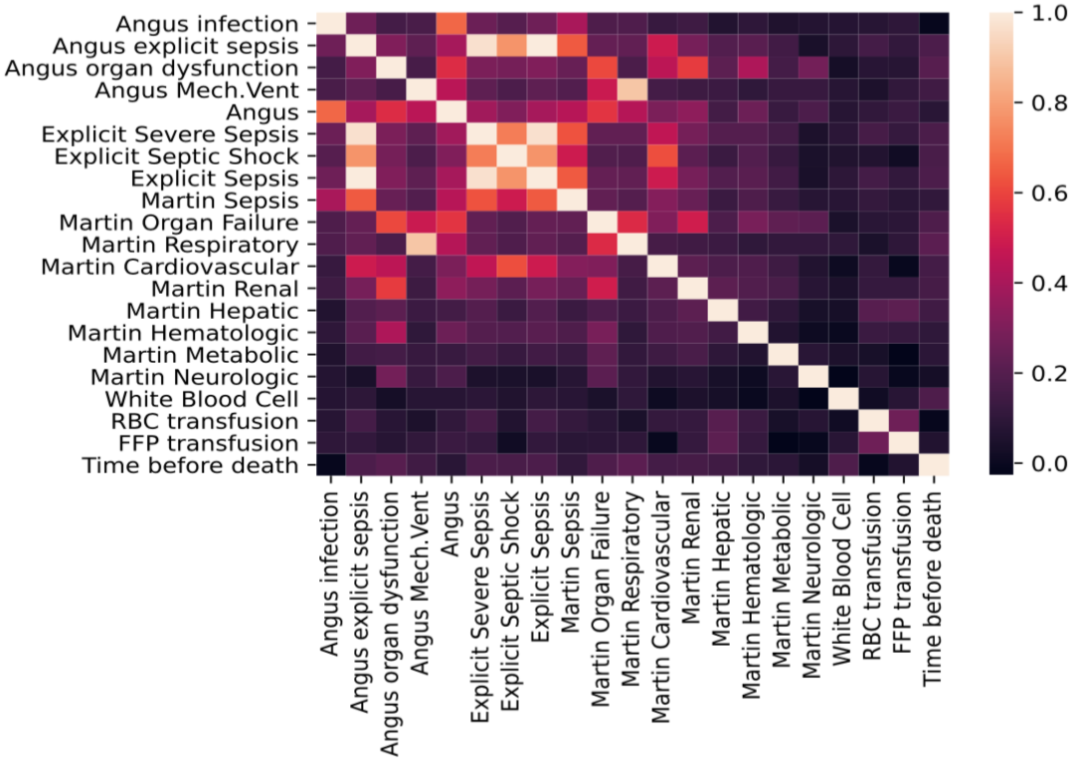

Figure 5.Feature. Feature correlation results (heatmaps) for sepsis scores

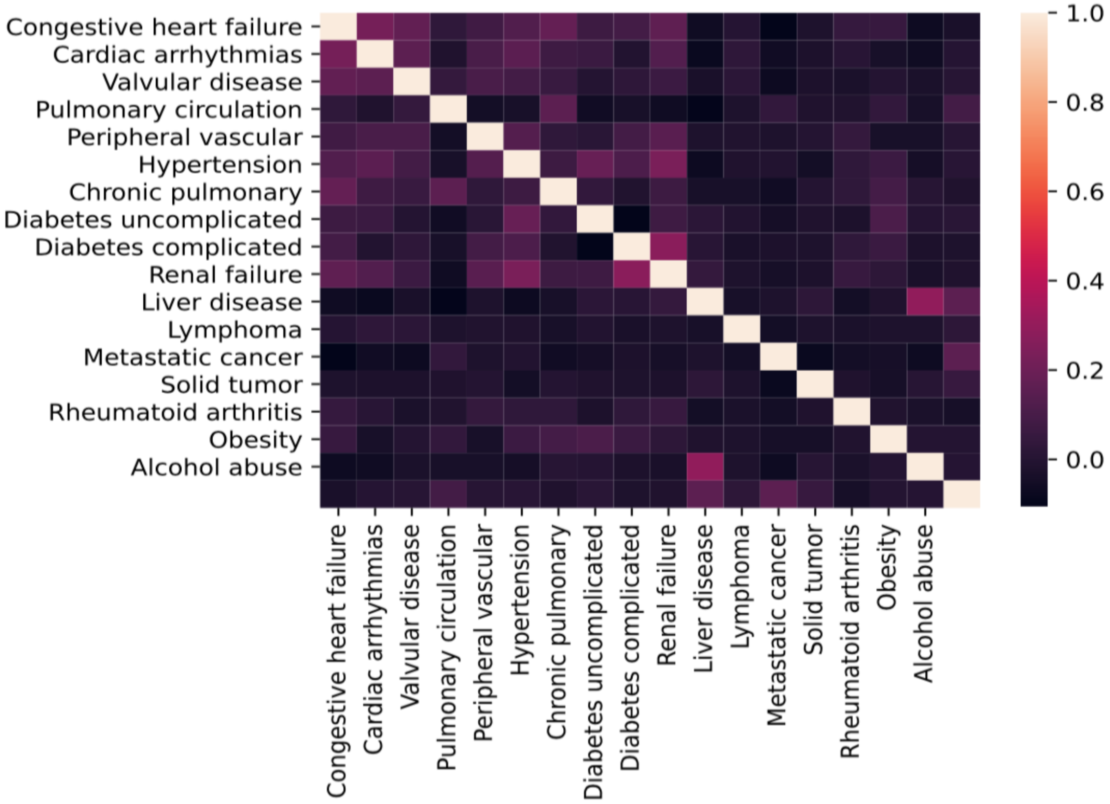

Figure 6.Feature correlation results (heatmaps) for comorbidities

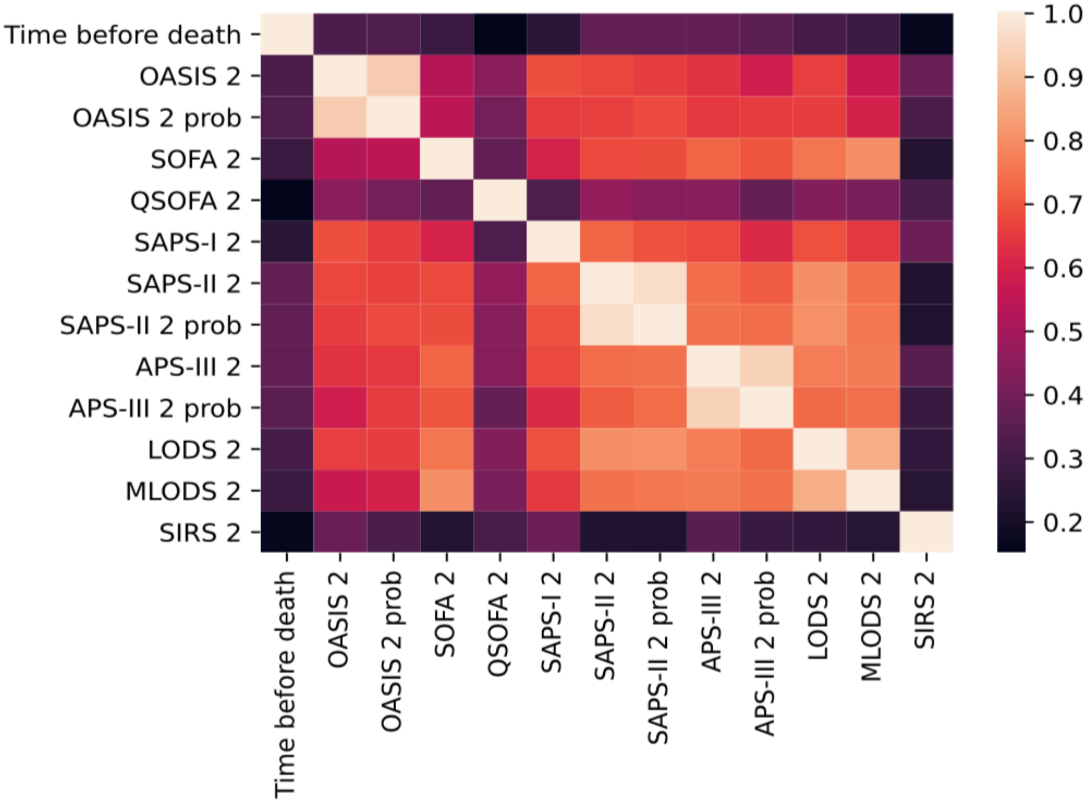

Figure 7. Feature correlation results (heatmaps) for ICU severity scores.

Abbreviations: APS= Acute Physiology Score, LODS= Logistic Organ Dysfunction Score, MLODS=Multiple LODS, OASIS= Outcome and Assessment Information Set, SAPS= Simplified Acute Physiology Score, SIRS=Systemic Inflammatory Response Syndrome, SOFA=Sequential Organ Failure Assessment, QSOFA=Quick SOFA score

5.2.2 Classification of early and late mortality

The best ML model chosen by JADBIO to predict early mortality (task M1) was Random Forests training 500 trees with Deviance splitting criterion and minimum leaf size = 3. As expected the best performance had the dataset containing all groups (AUC=0.925), followed by Concepts (AUC=0.923) and Chart Events (0.917), whereas Input Events had the worst performance (AUC=0.781), as shown in *Figure 8*. To further evaluate the performance of individual features in task M1, a dataset with all features except Concepts was used. It is surprising that the combination of all data preserves its high predictive performance (AUC 0.931). This is probably attributed to the equally high predictive performance of Chart Events.

Regarding late mortality (task M2), the best ML model was again Random Forests training 500 trees with Deviance splitting criterion and minimum leaf size = 3. Nevertheless, the task of predicting late mortality was less efficient even with the holistic approach (AUC=0.82). Concepts in this case had inferior performance (AUC=0.783) (*Figure 9*), which is expected since known severity and organ failure scores are excellent only for predicting early mortality, but it is interesting that they had the best performance comparing with the other feature sets. This difference can also be attributed to the fact that an unknown number of patients in the "alive" (G3) group might in fact have the same mortality risk as in the patients in G2 group due to the limited time period that the database tracks mortality status. *Table 6* describes the detailed metrics of the performance for both tasks M1 and M2 using all features, whereas *Table 7* describes the winning algorithms for each group of features set together with their corresponding AUC. Another interesting finding is that Note Events (free text features) had almost the same AUC (0.762) as Chart Events (0.768) and Procedures (0.763). This signifies the need to treat textual information as having the same importance for the classification task as with "traditional" clinical features, at least in ML tasks with a convoluted class distribution.

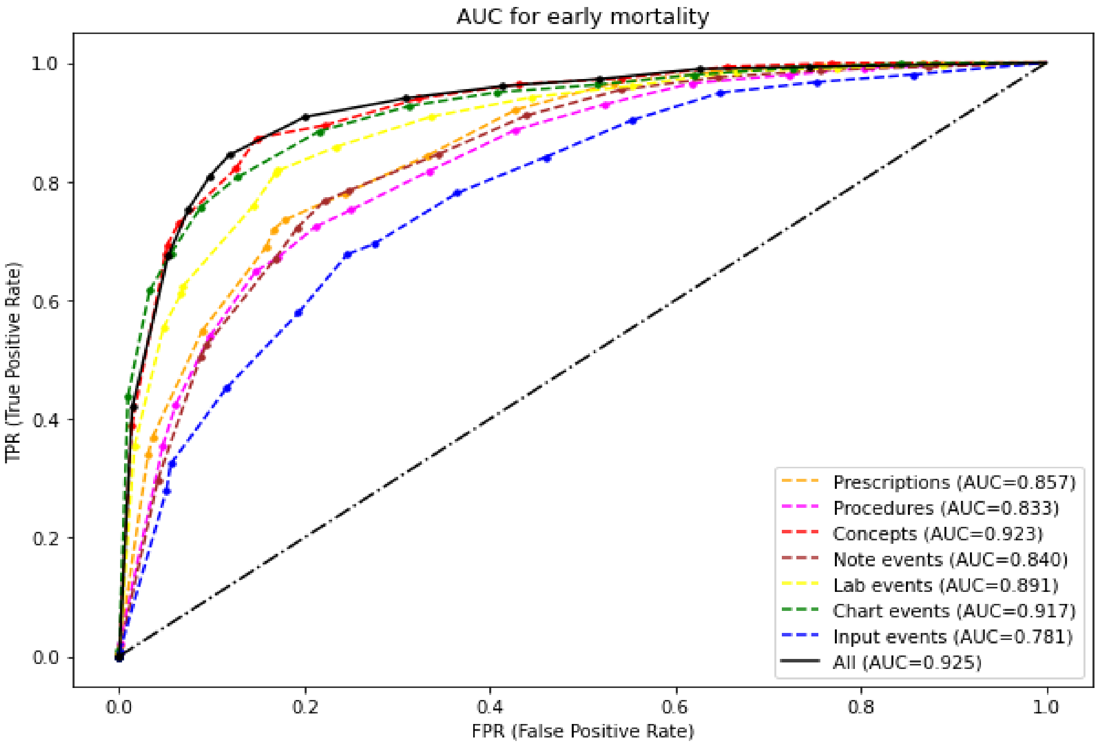

Figure 8. AUC for early mortality in ICU patients with thrombosis (Random Forest classifier).

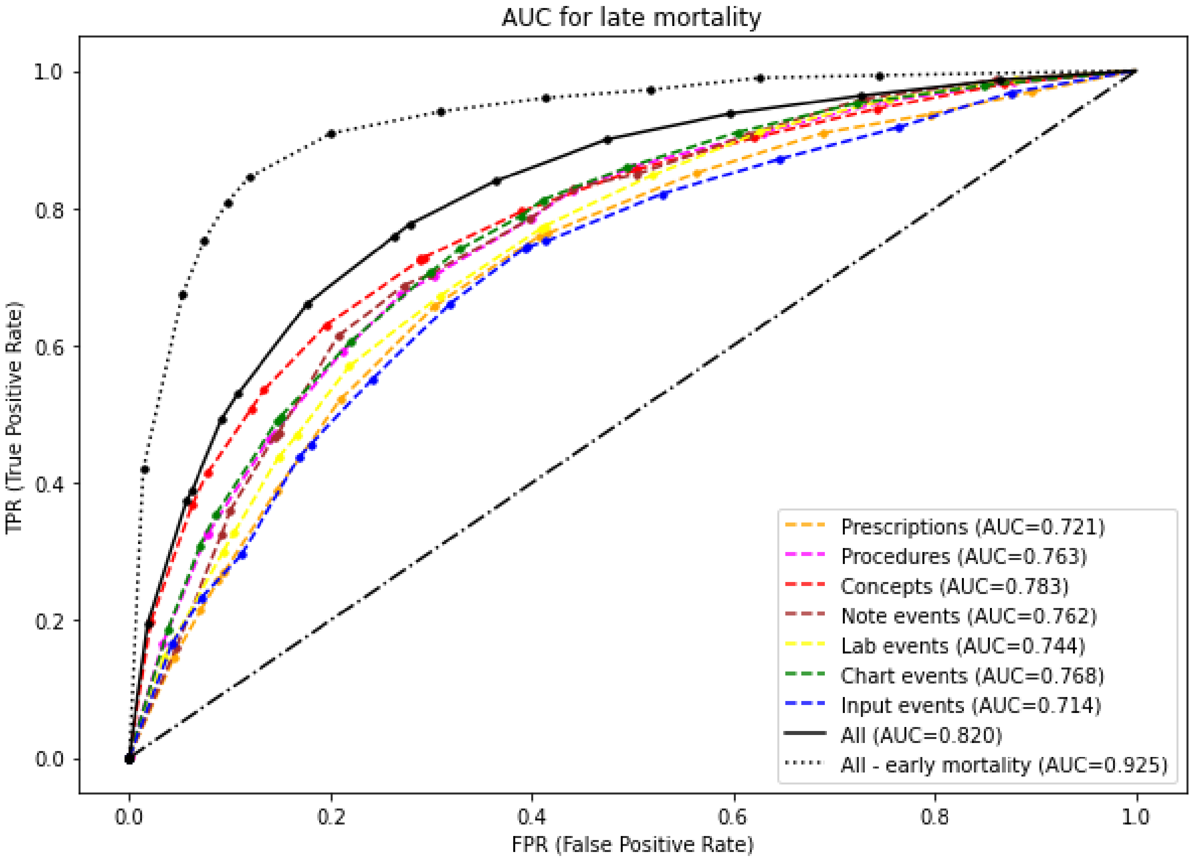

Figure 9. AUC for late mortality in ICU patients with thrombosis (Random Forest classifier).

Table 7. Detailed metrics of the performance for prediction of early and late mortality of ICU patients with thrombosis using all features.

	Early mortality	**Late mortality**
AUC [95% Confidence interval)	0.93 [0.91-0.95]	0.82 [0.79-0.84]
Accuracy	0.89	0.76
Balanced accuracy	0.81	0.74
F1 score	0.72	0.60
Precision	0.77	0.77
Sensitivity	0.67	0.49
Specificity	0.95	0.90

Table 8. Selected ML algorithms for each group of features.

	Early mortality		**Late mortality**	
	Feature selection	**Predictive algorithm**	**Feature selection**	**Predictive algorithm**
All	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.925]*	SES algorithm with hyper-parameters: maxK = 3, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.82]*
Chart Events	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.917]*	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.01	Ridge Logistic Regression with penalty hyper-parameter lambda = 0.1 *[AUC=0.768]*
Lab Events	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05 [AUC=0.744]	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.891]*	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 4 *[AUC=0.744]*
Procedures	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 4 *[AUC=0.833]*	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 2 *[AUC=0.763*]
Input Events	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 100 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.781]*	SES algorithm with hyper-parameters: maxK = 3, and alpha = 0.01	Ridge Logistic Regression with penalty hyper-parameter lambda = 10.0 *[AUC=0.714]*
Prescriptions	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 4 *[AUC=0.857]*	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 100 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.721]*
Note Events	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 100 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.840]*	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.01	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.762]*
Concepts	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 100 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.923]*	SES algorithm with hyper-parameters: maxK = 2, and alpha = 0.05	RF training 500 trees with Deviance splitting criterion and minimum leaf size = 3 *[AUC=0.783]*

Abbreviations : RF= Random Forest, SES=Statistically Equivalent Signature

5.2.3 Mortality prediction based on SMOTE and Random Forest

Figure 10 depicts the average ROC curves in the case of M1 (solid lines), where it is obvious that SMOTE oversampling (combined with the RF classifier) provides equal mean ROC results (0.91 in SMOTE and no-SMOTE case), something that was being expected due to the low imbalance ratio 1:3.744. Slightly better mean Precision-Recall (PR) scores are depicted in *Figure 11*. Since the imbalance ratio in the case of M2 is even lower (i.e., 1:1.595) it is expected that

SMOTE oversampling will achieve almost the same (or slightly worse) performance in comparison with the non-oversampling case. This is experimentally confirmed as it can be seen in *Figure 10* (dashed lines) where the mean ROC scores are 0.81 and 0.82 in the case of SMOTE and no-SMOTE, respectively, while the mean PR is the same for SMOTE and no-SMOTE as depicted in *Figure 12*.

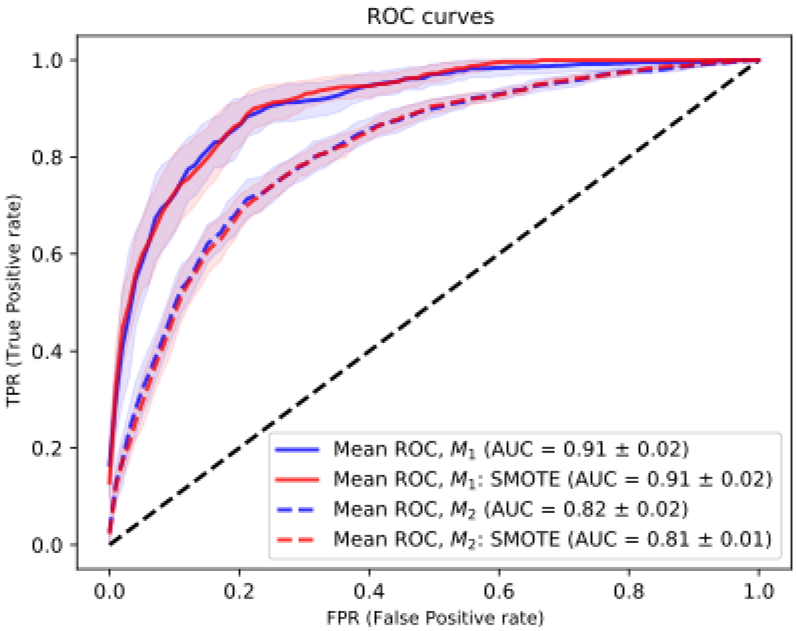

Figure 10. Average ROC curves for early (M1) and late (M2) mortality based on SMOTE and RF.

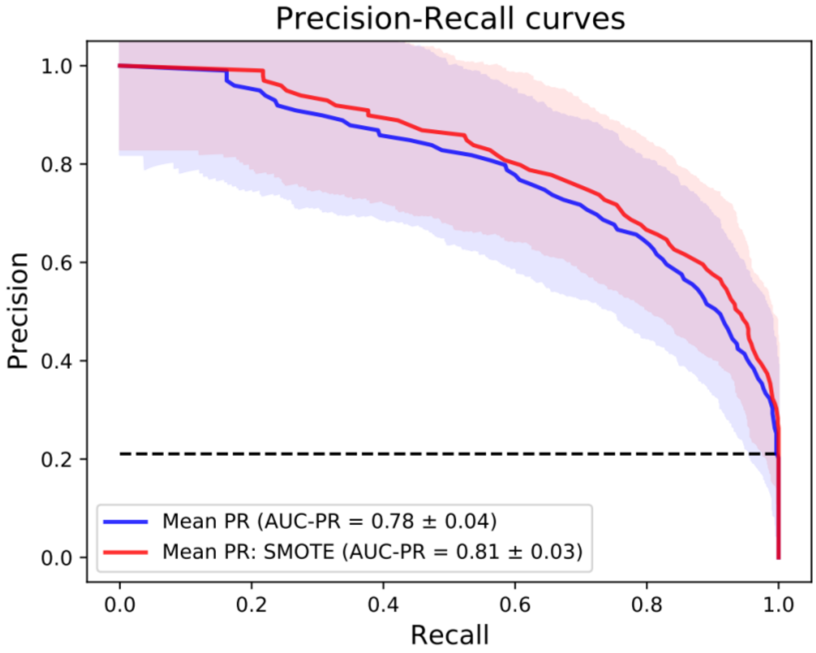

Figure 11.Average PR curves on early mortality (M1).

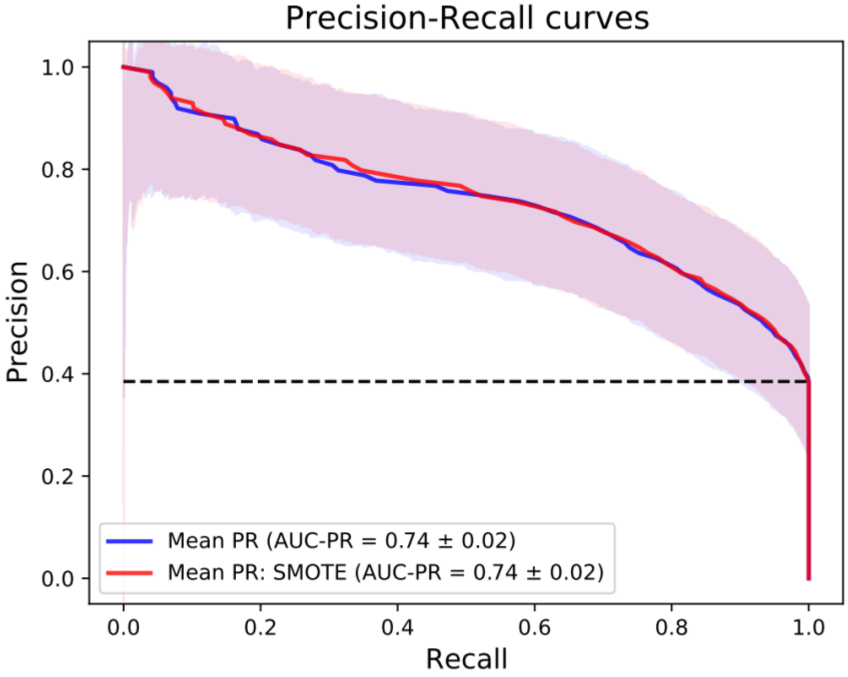

Figure 12. Average PR curves on late mortality (M2).

5.2.4 Feature Discriminative Analysis

In the next section the top selected features predictive of early and late mortality are shown, as well as their discriminative power for each group of features, that is medications (Prescriptions), charted patient data (ChartEvents), laboratory measurements for each given patient (LabEvents), transfusions and parenteral nutrition (InputEvents), performed procedures (Procedures) and known severity and sepsis scores (Concepts). For more details regarding the groups of features used in this study and algorithms used for feature selection, refer to *Table 4* and *Table 7* respectively. Moreover, how the prediction is enhanced by each selected feature when permuting all other features is reported (Table 9 and Table 10).

Most common features associated with mortality are shown in Figure 13, Figure 14, Table 9 and Table 10, where their permutation feature importance is shown in numbers. Since “NoteEvents” are not directly extracted features from the database, they do not have explicit clinical interpretation therefore they were not included in the figure and tables containing the most discriminatory features for each feature set.

Cancer and age at thrombosis were significant predictors in most of the analysis subgroups for early as well as for late mortality. Anticoagulation with warfarin in “All” and “Prescriptions” was another significant predictor for both M1 and M2. Selected features to predict M1 were features related to respiratory distress, renal failure, cardiovascular compromise, severity scores, certain medications, transfusions and laboratory indices. In more detail, respiratory distress was represented by blood gases (arterial pH, 1st day oxygen saturation), respiratory parameters of Martin sepsis score and respiratory rate (RR) in “All”, “Concepts” and “Chartevents” as well as mechanical ventilation and insertion of endotracheal tube in Procedures. Renal failure was indicated by blood urea nitrogen (BUN) in “All” and “ChartEvents”', urine output in “Concepts”, and creatinine in “LabEvents”. Cardiovascular compromise related-features were systolic (SBP) and 1st day diastolic blood pressure (DBP) in “Concepts”, extracorporeal circulation, cardiopulmonary resuscitation and infusion of vasopressors in “Procedures”, dopamine and norepinephrine administration in “Prescriptions”. From all severity scores, SAPS II appeared to significantly affect early mortality in "All", and "Concepts". GCS and mental status appeared as significant predictors in “ChartEvents”. Finally, well known significant laboratory indices (such as red cell distribution width or else RDW, platelets, white blood cells) were recognised in “LabEvents'' and "All" datasets.

Selected predictive features for late mortality were similarly associated with cardiovascular and renal failure, medications and laboratory indices. Renal failure was indicated by creatinine average, urine output, 1st day anion gap in “All” and “Concepts” and hemodialysis in “Procedures”. Cardiovascular compromise was represented by phenylephrine rate, blood pressure measurements and Creatine Phosphokinase (CPK) in “All” and “Chartevents” and extracorporeal circulation in “Procedures”. It is interesting that hydropneumothorax, a condition related to lungs,was a feature extracted from “NoteEvents”.

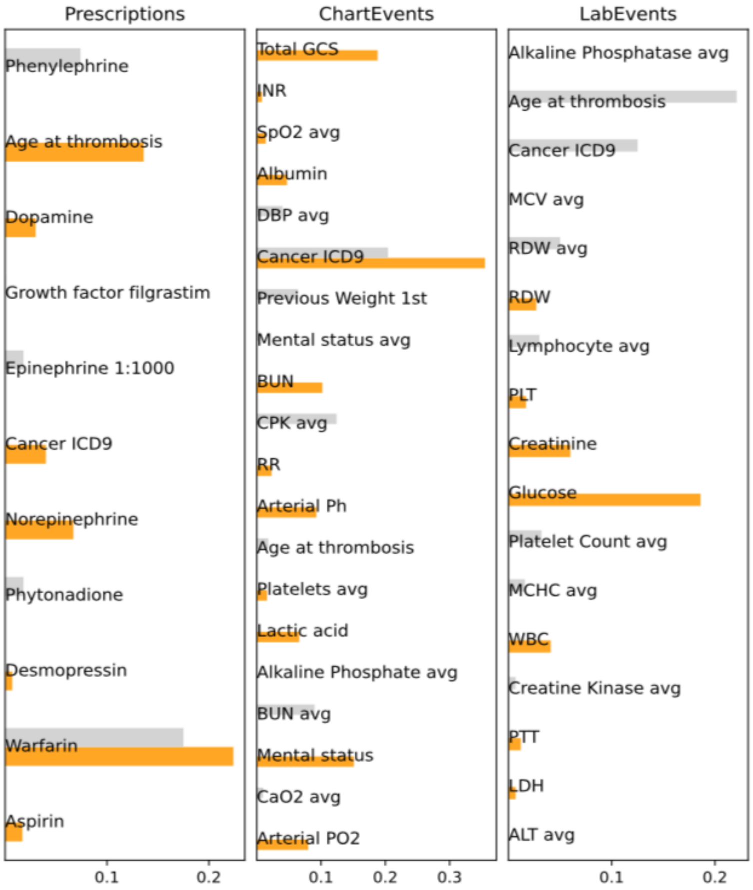

Figure 13. The discriminative power for the top features selected from JADBIO for Prescriptions, ChartEvents and LabEvents. Values represent the relative change of the AUC. Orange bars represent features for early mortality (M1) and Gray bars represent features for late mortality (M2).

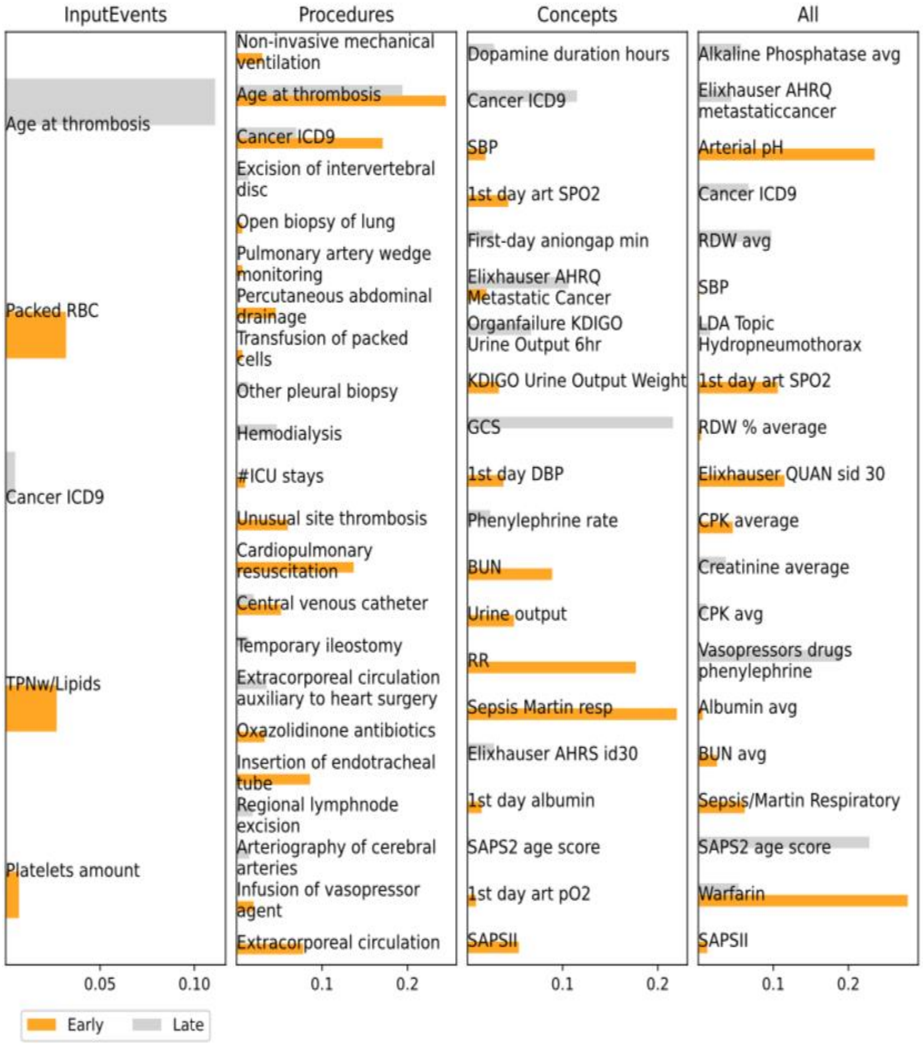

Figure 14. The discriminative power for the top features selected from JADBIO for InputEvents, Procedures, Concepts and All features. Values represent the relative change of the AUC. Orange bars represent features for early mortality (M1) and Gray bars represent features for late mortality (M2).

Abbreviations: ALT=Alanine Aminotransferase, Art pO2=Arterial Oxygen Partial Pressure, avg=average, BUN=Blood Urea Nitrogen, CaO2=Arterial Oxygen Content, CPK=Creatine Phospho-Kinase, DBP=Diastolic Blood Pressure, GCS=Glascow Coma Scale, INR=International Normalized Ratio, KDIGO=Kidney Disease Improving Global Outcome, LDH=Lactate Dehydrogenase, MCV=Mean Corpuscular Volume, PLT=Platelet, PTT=Partial Thromboplastin Time, RDW=Red Cell Distribution Width, RR=Respiratory Rate, SAPS=Simplified Acute Physiology Score, SBP=Systolic Blood Pressure, SpO2=Oxygen Saturation, TPN=Total Parenteral Nutrition, WBC=White Blood Cell.

Table 9. Selected most common features to predict early mortality and their permutation feature importance (number in brackets) that shows the increase in the prediction error of each selected feature when permuting all other features.

Features	Prescriptions	Chart events	Lab events	Input	Procedures	Concepts	All
F1	**Warfarin (0.224)**	**Cancer icd9 (0,355)**	Glucose (0.186)	Packed_RBC (0,032)	**Age_at_thrombosis (0.245)**	Sepsis Martin resp (0.22)	**Warfarin (0.279)**
F2	**Age at thrombosis (0,136)**	Total GCS (0,188)	Creatinine (0.06)	TPN_w/Lipids (0,027)	**Cancer_icd9 (0.171)**	SAPS II (0.102)	Arterial pH (0.235)
F3	Norepinephrine (0,067)	Mental status (0,151)	WBC (0.041)	PLT__amount (0,007)	Cardiopulmonary resuscitation, not_otherwise specified (0.137)	Glucose_avg (0.082)	Elixhauser-score-quan_sid30 (0.115)
F4	**Cancer icd9 (0,04)**	BUN (0.102)	RDW (0.027)		Insertion_of_endotracheal_tube (0.086)	**Cancer Icd9 (0.073)**	1st day blood-gas-art spo2 (0.106)
F5	Dopamine (0,03)	Arterial PO2 (0.08)	PLT (0.017)		Extracorporeal Circulation Auxiliary to_open_heart surgery (0.078)	Cookbook/WBC (0.069)	Sepsis/martin_respiratory (0.062)

Abbreviations: art (arterial), avg (average), BUN (blood urea nitrogen), DBP (diastolic blood pressure), Elixhauser_sid30 (depression), GCS (Glascow coma scale), MCHC (mean corpuscular hemoglobin concentration), PLT (platelets), RR (respiratory rate), RBC (red blood cells), RDW (red cell distribution width), SAPS II (Simplified acute physiology score II), TPN (total parenteral nutrition), WBC (white blood cells).

Table 10. Selected most common features to predict late mortality and their permutation feature importance that shows the increase in the prediction error of each selected features when permuting all other features.

Features	Prescription	Chart events	Lab events	Input	Procedures	Concepts	All
F1	**Warfarin (0.175)**	**Cancer_icd9 (0,204)**	**Age_at_thrombosis (0.221)**	**Age_at_thrombosis (0,111)**	**Age_at_thrombosis (0,194)**	GCS (0.216)	SAPS_age_score (0,228)
F2	Phenylephrine(0,074)	Creatine Kinase_avg (0,124)	**Cancer_icd9 (0.125)**	**Cancer_icd9 (0,005)**	**Cancer_icd9 (0,07)**	**Cancer_icd9 (0.115)**	Vasopressors_drugs__phenylephrine (0,19)
F3	Phytonadione (0,018)	BUN_avg (0,09)	**RDW_avg (0.05)**		Hemodialysis (0,047)	Elixhauser-ahrq-v37-no-drg-all-icd__metastatic_cancer (0.106)	**RDW_avg (0,097)**
F4	Epinephrine 1:1000 (0,018)	Previous Weigh_first (0,064)	PLT Count_avg (0.032)		Extracorporeal_circulation_auxiliary_to_open_heart_surgery (0,035)	Organfailure/kdigo-urineoutput_6hr (0.067)	**Cancer_icd9 (0,067)**
F5	Growth factor filgrastim (0,001)	DBP_avg (0,04)	Lymphocyte_avg (0.03)		Central_venous_catheter_placement_with_guidance (0,02)	Elixhauser-score-ahrq__elixhauser_sid30 (0.028)	Alkaline Phosphatase avg (0,057)
F6	Glipizide (0,0004)	**Age_at_thrombosis (0,018)**	MCHC_avg (0.016)		Regional_lymph_node_excision (0,019)	Dopamine-duration_hours (0.028)	**Antithrombotic_drugs__warfarin (0,054)**

Abbreviations: Comorbidity/Elixhauser-score-ahrq__elixhauser_sid30 (depression), DBP (diastolic blood pressure), GCS (Glascow coma scale), MCHC (mean corpuscular hemoglobin concentration), RDW (red cell distribution width), SAPS (Simplified acute physiology score).

6 DISCUSSION

The focus of the present study was to explore and assess whether a machine learning approach can be applied and further contribute in the prediction of two important research questions, that is the prediction of VTE in patients discharged from ICU as well as prediction of mortality in ICU patients with VTE. Towards this direction, a big-data driven research approach has been applied. One of the main goals of this work was to locate clinically meaningful predictive features that could probably contribute in building new models or refining existing scores to improve ICU and post-discharge survival rates. Moreover, the effect of balancing techniques on the final performance of the model was studied. Part of this study has been published at BIBE 2020 conference[98].

In this section, the initial aim of the study will be readdressed and results will be discussed. First, a summary of what has been published so far and most important problems with these studies will be reported. Then the inherent problems of the dataset chosen for this study will be presented and what would be done to improve the quality of the data. The use of machine learning algorithms to predict the above-mentioned research questions and to select clinical features will be analysed, as well as interpretation of the results will be given. Finally, comparison with published studies, limitations of this study and future work will be described.

The use of ML in prediction of DVT has been lagged so far. Only a few studies have recently tried to predict thrombosis in cancer patients[53, 70, 71] with questionable results and several methodological issues, as discussed earlier in the introduction. Several scores exist for prediction of VTE risk in hospitalized patients such as IMPROVE[99] and Padua[100] but no validated VTE risk assessment score exists specifically designed for critically-ill patients. Recently an ICU-VTE score derived from a retrospective analysis of a large number of patients using multivariate analysis has been published but not yet externally validated [101]. Similarly, Nafee et al, have used ensemble learning algorithms to predict VTE in critically-ill patients and report that their model outperforms classic statistical IMPROVE score[103]. Another important problem in thrombosis prediction, is that diagnosis of VTE can be frequently difficult or even masked, since VTE can be asymptomatic, or remain undiagnosed in a case of a sudden death following discharge from ICU

[104]. The risk of VTE after discharge from the hospital can be persistent for a significant time of period, depending on the reason of hospitalization and has been reported to be extended up to 3 months [105]. For that reason, this study focused on prediction of VTE readmission up to 90 days after discharge from ICU in cancer patients.

Prediction of early and late mortality in ICU patients has been also a central challenge in the area of medical informatics. Mortality is a major end point in epidemiological and interventional studies in the ICU, although somewhat debatable[106]. Published studies so far, focus mainly on the prediction of in-hospital mortality and use either a limited pre-selected number of features [22, 107, 108] or explore the feature space with a small range of ML algorithms (i.e., Logistic Regression[132] , SVM [70], Artificial Neural Networks[109], Decision Trees[110]). Even when more generic approaches are used, it is questionable whether proper ML guidelines for overfitting prevention and accurate performance metrics reporting are recorded. Moreover, most of the studies focus on in-hospital ICU mortality, irrespective of the primary patient diagnosis and their comorbidities. Since the initial diagnosis of the patient and the reason for ICU admittance could significantly affect overall survival, it would be interesting to study different disease outcomes and try to identify specific disease-related clinical features that have prognostic significance. Even more importantly, it is necessary to predict post-discharge mortality which is an even more difficult task, since patients admitted to ICU usually suffer from a high comorbidity burden. Few studies have focused on the prediction of late mortality in ICU patients such as [22].

Another important limitation of published predictive scores on ICU mortality, is that most of them are based on data recorded on the first day of ICU admission or the worst recorded value [111, 112 ,113]. MIMIC-III database gives the opportunity to study post-hospitalization mortality and long-term survival, since it collects information even months after discharge as well as readmissions to ICU. Using ML, complex algorithms can be trained on time-series data of a large population. This approach could possibly identify patterns or therapeutic interventions that could improve long-term survival in critically-ill patients[114]. It has been also shown that medical scores lose their performance over the time due to changes in medical practices[115]. Artificial intelligence could be promising in this direction, as it could be used as an adjunct tool that captures “live” data from electronic health records and assists clinicians to handle big-data collected over the process of clinical care [116].

The most difficult part of this study was to initially collect the appropriate database. To successfully predict thrombosis, a wide range of clinical, laboratory and genetic data are needed as well as clinical notes from prior medical history, comorbidities, known provoking factors such as recent surgeries and cardiovascular risk factors, prescriptions such as hormonal replacement and family history of thrombosis. This is due to the complex nature of the disease. In that direction, this study tried to include as many features as possible, that is a very wide range of raw demographic, clinical and laboratory measurements, widely accepted severity, comorbidity and organ failure scores (presented here as Concepts) [18] [19] [20] medications, procedures, transfusions as well as information coming from free-text notes. Unfortunately, genetic data, personal and family history of thrombosis are missing in the MIMIC III database. Notwithstanding, this holistic approach creates a feature space that except from the "curse of dimensionality", suffers from all known problems of real-world clinical data; imbalanced classes, many missing values[117] and co-dependencies between different features.

Extracting information as keywords or tokens from free medical notes is expected to be quite challenging for several reasons. First of all, the structure of notes is not homogeneous between different patients with different disease entities and between various caregivers of different medical specializations. Clinicians use very frequently different abbreviations or terms for the same diseases, either known medical abbreviations for example cerebrovascular accident (CVA) or more colloquial such as stroke. Sometimes clinicians describe the absence of symptoms, e.g. "without dyspnea". In this case the algorithm would probably keep the term "dyspnea" discarding "without", an action that could totally change the meaning of the sentence. Another example is the phrase "family history of a disease, e.g. stroke". In this case, the algorithm could keep only the word "stroke" without considering the context of the family history which is also important but from a different point of view. It would be probably interesting to focus on discharge notes, since these summarize the medical problems and the medications given to the patients, information quite important especially for post-discharge outcome.

Therefore, to incorporate unstructured free medical text, in the feature dataset, it was necessary to convert it to a numerical format. SABER [79] was used as a tool for information extraction and more specifically a pretrained model called DO which contains several entities related to signs, symptoms and diseases. SABER used free notes as input and extracted entities or tokens (that belong to the ontology) for each patient, and the final output of this pipeline was a vector of 50 topic models or clusters for each patient. One limitation of this method is that search is based on general medical terms and not focused on thrombosis-related entities, as discussed later in this section. It also produced non-interpretable features but with significant predictive abilities in ML algorithms, as shown in the case of early mortality prediction where NoteEvents had the same predictive performance as Prescriptions. Moreover, ideally each topic model contains tokens akin to specific groups of medical conditions or symptoms, for example cardiovascular diseases. A first attempt of predicting VTE in patients from electronic health records has been published from Sabra et al [133] using combined semantic and sentiment analysis.

To construct a robust model, training of two different ML strategies have been employed. The first, is an automated ML approach based on JADBIO platform, that has been widely tested in biomedical data and follows all good practices for analysis and efficiency reporting [89]. This platform uses a variety of ML algorithms such as RFs and SVMs that according to the authors have strong mechanisms to shield against overfitting. During the experiments, Random Forests were consistently found to be the winning algorithm. Besides that, JADBIO can produce "interpretable" models that can be intuitively explored and explained by physicians[89], as confirmed by this study. For feature selection, statistically equivalent signature (SES) algorithm, inspired by the principles of constrained-based learning of Bayesian networks, were consistently found to be superior against the feature selection method LASSO. All extracted features were clinically meaningful since older age, cancer, respiratory, cardiovascular, renal disease, vasopressor support and mechanical ventilation are well established clinical predictors of ICU mortality [22]. Similarly, with Ho et al.[22] sex was not found to be a predictor of ICU mortality. Moreover, individual feature analysis confirmed that warfarin [118] RDW [119], red blood cell transfusions [120] and blood urea nitrogen [121] are significant predictors of early and possibly long-term mortality. RDW has been shown to play a significant negative predictive role in ICU early mortality through deregulation of erythropoiesis from inflammatory cytokines and oxidative stress [122]. It has also been reported to be an independent risk factor for cardiovascular diseases, dyslipidemia, diabetes, renal and liver diseases. Surprisingly, high RDW has been shown to correlate with cancer stage irrespective of comorbidities and with early mortality in VTE patients. For all these reasons, it is not paradoxical that RDW could be an easily applicable, new biomarker, useful not only for the prediction of early but possibly of late mortality.

The second ML approach corresponds to a class balancing method which is combined with a Random Forest (RF) classifier, towards examining the fact that the results obtained from JADBIO are not affected by any class imbalance behavior. Typically, class imbalanced datasets constitute a common problem in medical informatics, which might lead to degraded performance depending on the type/number of data, features etc., and thus an additional analysis should be performed in order to tackle this issue. JADBIO addresses imbalanced classes through stratified cross-validation and diversified class weights during SVM learning, and thus an additional ML pipeline was implemented by adopting SMOTE method which is considered a state-of-the-art class balancing algorithm within the oversampling techniques framework[123].

In the case of the first ML task, i.e., the prediction of VTE-associated readmission in cancer ICU hospitalized patients, up to 90 days after discharge, all of the examined ML algorithms (Ridge logistic regression, SVM, DT, RF) failed to accurately provide a high predictive performance. Efforts to improve imbalance ratio via SMOTE and reduce dimensionality of the dataset were unsuccessful and they did not add any drastic change in the overall predictive accuracy. This is probably due to the nature of the dataset, since the number of positive cases was really low comparing with the number of negative cases. Even if the algorithm failed to predict efficiently thrombosis, features selected such as multiorgan failure and sepsis scores, insertion of endotracheal tube, and red blood cell transfusions are meaningful in the context of thrombosis[124]. Similarly, in the case of the second task, i.e., the prediction of early and late mortality in thrombosis ICU patients, SMOTE did not significantly change the predictive accuracy as compared to not applying SMOTE during the ML procedure. Nevertheless, features selected like red cell transfusions[127] and endotracheal intubation[128] are well known correlated with ICU mortality. Further work is needed towards this direction, such as balancing the classes through other resampling techniques that might be more successful in cases of extremely high-class imbalanced ratios e.g. generative adversarial networks [129].

There is a growing number of studies that apply ML to optimize early prediction of various clinical tasks but unfortunately direct comparison between them is not yet feasible due to different methodological approaches, different definitions for the same parameters and heterogeneous population [130]. In the current study, a thorough analysis of early and late mortality ML-based prediction of patients diagnosed with venous thromboembolism is provided. More specifically, a multi-feature analysis was performed based on data obtained from MIMIC-III database, and textual input via the NoteEvents was added to enrich the features pool. A bias-free prediction accuracy is also provided based on autoML pipeline. The main outcome of the prior analysis is that the concepts, that is meta-features provided in MIMIC III database, seem to be indicative for accurate early mortality prediction. However, concepts seem to have an inferior performance for late mortality prediction, where additional clinical features need to be added. It is quite interesting though, that concepts retain the highest performance comparing with all other feature sets, making them promising for future research. It is important to mention that late mortality prediction in light of ML binary classification, is far more challenging than early prediction, and this seems to be an inherent problem of MIMIC-III database.

In order to explore the predictive role of various feature groups the analysis was scaled in 8 distinct groups. For the early mortality, combining all features had the best performance followed by Concepts and ChartEvents, whereas regarding late mortality, the prediction task was less efficient even with the holistic approach. Prediction of late mortality is expected to be more challenging since follow up of patients is suboptimal and presence of other unpredictable factors can alter outcome. Another interesting finding is that NoteEvents (free text features) had almost the same predictive performance as ChartEvents and Procedures. This signifies the need to treat textual information as having the same importance for the classification task as with “traditional” clinical features, at least in ML tasks with a convoluted class distribution. InputEvents had the worst prognostic value for both early and late mortality. This could be attributed to the redundancy of the two systems used that could possibly affect the prognostic significance of these procedures since grouping of the two different systems was not performed. This redundancy resulted in a significant number of missing values since patients recorded with the one system were not recorded with the other. Overall, it seems that to predict late mortality the use of the maximum number of features contributed to improved predictive capability.

Concepts contain valuable information for predicting early mortality, reaching the same efficiency as the complete feature space, with a high AUC (0.923). Nevertheless, when predicting late mortality, information from all other groups can significantly increase the AUC from 0.783 to 0.82. Concepts in this case had inferior performance, which is expected since known severity and organ failure scores were originated only for predicting early mortality. As a comparison, one of the best existing studies in 442,692 patients for predicting 90-day mortality had AUC of 0.86 by leveraging 5,695 features[132].The model developed in the present study outperforms Cugno et al [108] in prediction of early mortality (AUC 0.92 vs 0.77). Also, in a recent review [131] of 43 mortality prediction models for critically-ill patients the lowest discrimination AUC was 0.72 and the highest 0.91. From these, the only one that used a multi-feature approach [132] had an AUC of 0.86 for 6-month mortality and 0.88 for 12-month mortality. Regarding ICU scores, Fuchs et al. [37] report AUC of 0.826, 0.836, and 0.788 for SAPS II, APACHE II, and SOFA scales, respectively, for predicting ICU mortality, and 0.708, 0.709, and 0.661 for SAPS II, APACHE II, and SOFA, respectively, for post-ICU prognosis. Therefore, there is a need for more precise and reliable tools for estimating long-term survival of the VTE patients successfully discharged from ICU. It would be probably interesting to focus on discharge data, as well as having a close monitoring after discharge which is probably impractical, since many patients are lost during follow-up.

The correlation study confirmed the hypothesis that different sepsis and comorbidity scores convey different types of information [21]. Regarding sepsis, it is interesting that there is a quite good correlation between the two sepsis scores (Angus and Martin), whereas surprisingly white blood cells, blood components transfusion and time before death do not seem to correlate well with sepsis. As expected in comorbidities scores, a moderate correlation between liver disease and alcohol abuse, renal failure with hypertension, diabetes, and hypertension are shown.

Finally, a strong correlation is observed between various severity and organ failure scores, although none of these scores showed a strong correlation with time before death.

Some limitations of this study should be considered. First, the study was retrospective from a single US medical center from the previous decade. Since the data were collected in the past, it is possible that many medical practices have changed over time, such as the case of warfarin. Second, the selection of the patients with thrombosis was based solely on ICD-9 codes[74] and DRG codes[75]. This could include some false negative and false positive cases, since confirmation by imaging studies was not feasible. Ideally the identification of patients subsequently developing thrombosis would be through imaging studies, which are not provided in this database. Third, no external validation of our results has been performed, since the primary goal of the study was to initially explore the feasibility of various ML approaches in prognostication of such tasks and not to provide a new score. Finally, a more focused approach of natural language processing such as Semantic Extraction and Sentiment Assessment of Risk Factors (SESARF) could be more effective and VTE oriented[133]. SESARF framework uses an algorithm to extract risk factors based on an expert approved list, semantic enrichment through a continuously updating dictionary, calculations of risk factors weight, sentiment assessment and development of a scoring model and finally prediction through classification with support vector machines (SVM). Another alternative approach would be instead of extracting specific ontology-based entities, to use direct language embeddings [134] .

One of the primary goals of future work is external validation of this prognostic model using eICU Collaborative Research Database, which is a larger and more recent database from different US hospitals [135]. Other future directions, include to focus on features extracted on the day of discharge to predict early readmission to ICU for patients with VTE, and early readmission to ICU of cancer patients. Finally, using LSTM for importing time series data in the model [114] and deep learning models to increase predictive performance in the long-term prognosis and possibly predict length of stay[136].

7 CONCLUSION

This study explored the application of machine learning in the prediction of two important clinical questions, the risk of ICU readmission in cancer patients due to thrombosis and early as well as late mortality of ICU hospitalized patients with VTE. It is important to accurately predict these two problems, since prompt recognition of thrombosis or mortality risk could re-orientate medical clinical practices (e.g. extended anticoagulation in patients with high risk of thrombosis) and help clinicians to reasonably allocate health resources in ICUs, which are extremely restricted, especially in the era of COVID-19 pandemic. Currently, no universally accepted medical scores exist to assist clinicians in decision making, since they have modest performance, limited generalizability, low objectivity and limited interoperability, as discussed above.

A big-data driven research approach was used as well as stratification over the different group of features. The study was performed in two retrospective cohort populations derived from an open access MIMIC III database. Prediction of VTE readmission within 90 days in cancer ICU patients was not feasible with either ML approaches, probably due to the extremely high imbalance ratio of the dataset, as well as the inappropriate follow up of patients attributable to the retrospective nature of the study, and that readmission was based solely on ICD9 codes and not on imaging studies. Moreover, the effect of class balancing techniques was examined to overcome the problem of high imbalance ratio, a frequent problem that real-life medical datasets suffer of, but without any favorable results. Nevertheless, it is quite interesting that the most important selected features in this case, were parameters of respiratory instability (such as endotracheal tube insertion), packed red blood cell transfusions, renal and hepatic dysfunction and MCH, an index of red blood cells.

On the other hand, early mortality in critically-ill patients with VTE can be easily predicted by Random Forests classifier, which is robust and efficient when dealing with complex data. As expected by combining all features (N=1,471) resulted in the highest efficiency, followed by concepts, that include organ failure scores and charted documents such as vital signs. Regarding the rest of the feature group it is noteworthy that textual information had comparable efficiency with medications, so further attention is needed when handling with medical notes. Important clinical predictors of early mortality are parameters of respiratory distress, cardiovascular compromise, renal failure, red cell transfusions, hematological parameters (white blood cells, RDW, platelets) and organ failure scores (SAPS II, CGS). Prediction of late mortality is slightly inferior but acceptable, due to the complexity of this task, the inherent problems of the database and the confounding comorbidities. Similarly, with early mortality, important clinical predictors are renal failure, cardiovascular compromise, organ failure scores and hematological parameters (RDW, platelets).

The herein research could be used as a proof of concept study that could be further validated in prospective or more recent databases. Inclusion of more features such as genetic information, personal and family history would be ideal and would probably improve predictive performance. The results of this study are promising and most importantly explainable, since the algorithm was able to select predictive features that were clinically meaningful, as already mentioned. For example, age, cancer, warfarin treatment, sepsis and severity scores, vasopressor support as well as RDW and platelet number were significant predictors for both early and late mortality. Sepsis and severity scores exert a modest performance in prediction of late mortality as expected, since these scores were initially derived for prediction of early mortality. Red blood cell transfusions are a negative predictor of early mortality as already known. RDW could be a candidate rediscovered easily applicable biomarker, since it is known that correlates with early ICU mortality, early mortality from VTE, as well as with other comorbidities, so it would be probably interesting to introduce it in current prognostic scores. There is a need for more precise and reliable tools in order to estimate late mortality in VTE patients successfully discharged from the ICU and risk of VTE in cancer patients discharged from ICU, since this knowledge could alter the clinical decisions and therapeutic interventions of medical practitioners. Implementing deep learning algorithms for these complex prediction tasks could probably improve performance of the model. Finally, although direct comparison of the proposed framework with known Risk Assessment Models or published data is not possible due to heterogeneous datasets, different study design, even various definitions of mortality, the results of this study lie on the top of existing classification performance for these ML tasks.

8 ABBREVIATIONS

AHPS	Algorithm and Hyper-Parameter Space
AI	Artificial intelligence
ALT	Alanine Aminotransferase
ANN	Artificial neural networks
APACHE	Acute Physiology and Chronic Health Evaluation
APS	Acute Physiology Score
Art pO2	Arterial Oxygen Partial Pressure
AUC	area under the curve
AutoML	automated machine learning
Avg	average
BBC-CV	Bootstrap Bias Corrected Cross-Validation
BUN	blood urea nitrogen
CaO2	Arterial Oxygen Content
CCS	Clinical Classification Software
CCU	Coronary Care Unit
CPK	Creatine Phospho-Kinase
CSRU	Cardiac Surgery Recovery Unit
CVA	cerebrovascular accident
DBP	diastolic blood pressure
DNR	do not resuscitate code
DO	disease ontology
DRG	diagnosis-related groups
DVT	deep vein thrombosis
EHR	electronic health records
FVL	factor V Leiden
GCS	Glasgow Coma Scale
HIPAA	Health Insurance Portability and Accountability Act
ICD-9	International Classification of Diseases 9th edition
ICU	intensive care units

INR	International Normalized Ratio
K-NN	K-nearest neighbor
KDIGO	Kidney Disease Improving Global Outcome
LDA	Latent Dirichlet Allocation
LDH	Lactate Dehydrogenase
LODS	Logistic Organ Dysfunction Score
LSTM	Long Short-Term Memory
MCH	mean corpuscular hemoglobin
MCHC	mean corpuscular hemoglobin concentration
MCV	Mean Corpuscular Volume
MICU	Medical Intensive Care Unit
ML	machine learning
NLP	natural language processing
PCA	Principal Component Analysis
PE	pulmonary embolism
PESI	Pulmonary Embolism Severity Index
PLT	Platelet
PR	Precision-Recall
PTT	Partial Thromboplastin Time
RAMs	risk assessment models
RBC	red blood cells
RBF	radial basis function
RDW	Red cell distribution width
RF	Random Forests
RF	Random Forests
RO	random optimization
ROC	Receiver operating curves
RR	respiratory rate
SAPS	Simplified Acute Physiology Score
SBP	systolic blood pressure
SESARF	Semantic Extraction and Sentiment Assessment of Risk Factors

SICU	Surgical Intensive Care Unit
SIRS	Systemic Inflammatory Response Syndrome
SMOTE	Synthetic Minority Oversampling Technique
SOFA	Sequential Organ Failure Assessment
SpO2	Oxygen Saturation
SVM	Support Vector Machines
TPN	Total Parenteral Nutrition
TSICU	Trauma Surgical Intensive Care Unit
VTE	venous thromboembolism
WBC	White Blood Cell.

9 REFERENCES

[1] Heit JA. Epidemiology of venous thromboembolism. Nature reviews. Cardiology. 2015;12: 464–474. doi:10.1038/nrcardio.2015.83

[2] Schulman S, Ageno W, Konstantinides S. Venous thromboembolism: past, present and future. Thromb & Haemost 2017; 117: 1119-27.

[3] Goldhaber SZ, Bounameaux H. Pulmonary embolism and deep vein thrombosis. Lancet. 2012; 379: 1835-1846. https://doi.org/10.1016/S0140-6736(11)61904-1

[4] Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012; 380:2095–2128. doi: 10.1016/S0140-6736(12)61728-0.

[5] Robinson GV. Pulmonary embolism in hospital practice. BMJ. 2006;332(7534):156-160. doi:10.1136/bmj.332.7534.156

[6] Bahloul M, Chaari A, Kallel H, et al. Pulmonary embolism in intensive care unit: Predictive factors, clinical manifestations and outcome. Ann Thorac Med. 2010; 5: 97-103. doi:10.4103/1817-1737.62473

[7] Prandoni P, Falanga A, Piccioli A. Cancer and venous thromboembolism. Lancet Oncol. 2005;6:401-410. doi:10.1016/S1470-2045(05)70207-2

[8] Hirsch DR, Ingenito EP, Goldhaber SZ. Prevalence of deep venous thrombosis among patients in medical intensive care. JAMA. 1995; 274: 335–7.

[9] McKenzie SB, Williams JL, 2015. Clinical Laboratory Hematology. 3rd edition. Pearson.

[10] Moheimani F, Jackson D. Venous thromboembolism. Classification, risk factors, diagnosis and management. ISRN Hematology. 2011. doi:10.5402/2011/124610.

[11] Salwa K, Dickerman JD. Hereditary thrombophilia. Thrombosis J. 2006; 4. Doi:10.1186/1477-9560-4-15.

[12] Stevens SM, Woller SC, Bauer KA, et al. Guidance for the evaluation and treatment of hereditary and acquired thrombophilia. J Thromb. Thrombolysis 2016; 41:154-164.

[13] Razak NB, Jones G, Bhandari M, Berndt MC, Metharom P. Cancer-Associated Thrombosis: An Overview of Mechanisms, Risk Factors, and Treatment. Cancers. 2018;10. pii: E380. doi: 10.3390/cancers10100380.

[14] Al-Samkari H, Connors JM. Managing the competing risks of thrombosis, bleeding, and anticoagulation in patients with malignancy. Blood Adv. 2019; 3: 3770-3779. doi:10.1182/bloodadvances.2019000369

[15] Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. Blood 2008; 111: 4901-070.

[16] Gerotziafas GT, Taher A, Abdel-Razeq H, et al. A Predictive Score for Thrombosis Associated with Breast, Colorectal, Lung, or Ovarian Cancer: The Prospective COMPASS-Cancer-Associated Thrombosis Study. Oncologist. 2017; 22(10): 1222–1231. doi:10.1634/theoncologist.2016-0414.

[17] Greene T, Spyropoulos A, Chopra V, et al. Validation of risk assessment models of venous thromboembolism in hospitalized medical patients. Am J Med 2016; 129: 1001.e9-e18. http://dx.doi.org/10.1016/j.amjmed.2016.03.031

[18] Le Gall JR, Lemeshow S, and Saulnier, F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA. 1993; 270: 2957–2963.

[19] Knaus, WA, Wagner, DP, Draper, EA et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991; 100: 1619–1636.

[20] Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA. 2001; 286(14): 1754-1758. doi:10.1001/jama.286.14.1754

[21] Vincent, J., Moreno, R. Clinical review: Scoring systems in the critically ill. Crit Care. 2010; 14: 207. https://doi.org/10.1186/cc8204

[22] Ho KM, Knuiman M, Finn J, Webb SA. Estimating long-term survival of critically ill patients: the PREDICT model. PLoS One. 2008; 3(9): e3226. doi:10.1371/journal.pone.0003226

[23] Sánchez-Hurtado LA, Ángeles-Veléz A, Tejeda-Huezo BC, García-Cruz JC, Juárez-Cedillo T. Validation of a prognostic score for mortality in elderly patients admitted to Intensive Care Unit. Indian J Crit Care Med. 2016; 20: 695-700. doi:10.4103/0972-5229.195702

[24] Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: A compendium. Indian J Crit Care Med. 2014; 18: 220-228. doi:10.4103/0972-5229.130573

[25] Chen LM, Martin CM, Morrison TL, Sibbald WJ. Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. Crit Care Med. 1999 Sep;27(9):1999-2004. Doi: 10.1097/00003246-199909000-00046.

[26] Kafeza M, Shalhoub J, Salooja N, . A systematic review of clinical prediction scores for deep vein thrombosis. Phlebology. 2017; 32: 516-531. doi: 10.1177/0268355516678729.

[27] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259

[28] Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood). 2014;33(7):1163-1170. doi:10.1377/hlthaff.2014.0053

[29] Ay C., Dunkler D., Marosi C., Chiriac A.L., Vormittag R., Simanek R., Quehenberger P., Zielinski C., Pabinger I. Prediction of venous thromboembolism in cancer patients. Blood. 2010; 116:5377–5382. doi: 10.1182/blood-2010-02-270116.

[30] Verso M., Agnelli G., Barni S., Gasparini G., Labianca R. A modified Khorana risk assessment score for venous thromboembolism in cancer patients receiving chemotherapy: The Protecht score. Intern. Emerg. Med. 2012; 7: 291–292. doi: 10.1007/s11739-012-0784-y.

[31] Cella C.A., Di Minno G., Carlomagno C., Arcopinto M., Cerbone A.M., Matano E., Tufano A., Lordick F., De Simone B., Arturo C., et al. Preventing venous thromboembolism in ambulatory cancer patients: The ONKOTEV Study. Oncologist. 2017; 22: 601–608. doi: 10.1634/theoncologist.2016-0246.

[32] Rupa-Matysek J, Lembicz M, Rogowska EK, Gil L, Komarnicki M, Batura-Gabryel H. Evaluation of risk factors and assessment models for predicting venous thromboembolism in lung cancer patients. Med Oncol. 2018; 35: 63. doi: 10.1007/s12032-018-1120-9.

[33] Wicki J, Perrier A, Perneger TV, Bounameaux H, Junod AF. Predicting adverse outcome in patients with acute pulmonary embolism: a risk score. Thromb Haemost. 2000; 84: 548.

[34] Aujesky D, Obrosky DS, Stone RA, Auble TE, Perrier A, Cornuz J, Roy PM, Fine MJ Derivation and validation of a prognostic model for pulmonary embolism. Am J Respir Crit Care Med. 2005; 172: 1041.

[35] Jiménez D, Kopecna D, Tapson V, Briese B, Schreiber D, Lobo JL, Monreal M, Aujesky D, Sanchez O, Meyer G, Konstantinides S, Yusen RD, On Behalf Of The Protect Investigators. Derivation and validation of multimarker prognostication for normotensive patients with acute symptomatic pulmonary embolism. Am J Respir Crit Care Med. 2014; 189: 718.

[36] Johnson AE, Kramer AA, Clifford GD. A new severity of illness scale using a subset of Acute Physiology And Chronic Health Evaluation data elements shows comparable predictive accuracy. Crit Care Med. 2013;41(7):1711-1718. doi:10.1097/CCM.0b013e31828a24fe.

[37] Fuchs PA, Czech IJ, Krzych ŁJ. The Pros and Cons of the Prediction Game: The Never-ending Debate of Mortality in the Intensive Care Unit. Int J Environ Res Public Health. 2019;16(18):3394. doi:10.3390/ijerph16183394

[38] Jiang F, Jiang Y, Zhi H, Dong Y, Li H et, al. Artificial intelligence in healthcare: past, present and future. Stroke & Vasc Neurol. 2017; 2: e000101. Doi: 10.1136/svn-2017-000101.

[39] Sheskin, David J. (2003) Handbook of Parametric and Nonparametric Statistical Procedures. CRC Press. ISBN 1-58488-440-1

[40] Witten IH, Frank E. Data mining. Practical Machine Learning and Techniques. Morgan Kaufmann Publishers. 2nd edition (2005).

[41] https://en.wikipedia.org/wiki/Logistic_regression

[42] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[43] https://en.wikipedia.org/wiki/Bayes%27_theorem

[44] https://en.wikipedia.org/wiki/Linear_discriminant_analysis

[45] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[46] https://en.wikipedia.org/wiki/Decision_tree

[47] Podgorelec, V., Kokol, P., Stiglic, B. et al. Decision Trees: An Overview and Their Use in Medicine. Journal of Medical Systems 2002; 26: 445–463. https://doi.org/10.1023/A:1016409317640

[48] https://en.wikipedia.org/wiki/Support_vector_machine

[49] https://en.wikipedia.org/wiki/Principal_component_analysis

[50] Jaadi Z. A step by step guide in Principal Component Analysis. Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[51] https://en.wikipedia.org/wiki/Ensemble_learning

[52] https://en.wikipedia.org/wiki/Artificial_neural_network

[53] Willan J, Katz H, Keeling D. The use of artificial neural network analysis can improve the risk-stratification of patients presenting with suspected deep vein thrombosis. Br J Haematol. 2019;185(2):289-296. doi: 10.1111/bjh.15780. Epub 2019 Feb 6. PMID: 30727024

[54] https://en.wikipedia.org/wiki/Reinforcement_learning

[55] Bellman R. A Markovian Decision Process. Journal of Mathematics and Mechanics, 1957; 6: 679-684.

[56] Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. Academic pathology. 2019; 6, 2374289519873088. doi:10.1177/2374289519873088

[57] https://en.wikipedia.org/wiki/Deep_learning

[58] Bengio, Y. Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning. 2009; 2: 1–127.doi:10.1561/2200000006.

[59] https://en.wikipedia.org/wiki/Natural_language_processing

[60] Kotthoff L, Thornton C, Hoos H, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. JMLR. 2017; 18:1−5.

[61] Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and Robust Automated Machine Learning Advances in Neural Information Processing Systems 28 (NIPS 2015).

[62] Xanthopoulos I, Tsamardinos I, Christophides V, Simon E, Salinger A. Putting the Human Back in the AutoML Loop. Conference: Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference

[63] Truong A, Walters A, Goodsitt J, Hines K, Bruss CB, FarivarR. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), 2019: 1471-1479, doi: 10.1109/ICTAI.2019.00209.

[64] Spyropoulos A, McGinn T, Khorana AA. The use of weighted and scored risk assessment models for venous thromboembolism. Thromb & Haemost 2012; 108: 1072-76

[65] https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0290

[66] Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35

[67] Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

[68] https://healthcare.ai/machine-learning-versus-statistics-use/

[69] Hassanipour S, Ghaem H, Arab-Zozani M, Seif M, Fararouei M, Abdzadeh E, Sabetian G, Paydar S. Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: A systematic review and meta-analysis. Injury. 2019; 50: 244-250. doi: 10.1016/j.injury.2019.01.007.

[70] Ferroni, P., Zanzotto, F. M., Scarpato N., Riondino S., Nanni U., Roselli M., Guadagni F. Risk Assessment for venous thromboembolism in chemotherapy-treated ambulatory cancer patients: A machine learning approach. Med. Decis. Mak. 2017; 37: 234–242.

[71] Z. Qatawneh, M. Alshraideh, N. Almasri, L. Tahat, A. Awidi, 2017. Clinical decision support system for venous thromboembolism risk classification. Applied Comput. Informat. DOI: 10.1016/j.aci.2017.09.003

[72] Caprini JA, Arcelus JI, Hasty JH, Tamhane AC, Fabrega F. "Clinical assessment of venous thromboembolic risk in surgical patients". Semin Thromb Hemost. 1991; 17 Suppl 3: 304–12.

[73] Wells PS, Anderson DR, Bormaniis J et al. Value of assessment of pretest probability of deep vein thrombosis in clinical management. Lancet. 1997; 350: 1795-8.

[74] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

[75] Busser R, Geisser A, Quentin W, Willey M. Diagnosis related groups in Europe. World Health Organization. McGraw-Hill 2011. https://www.euro.who.int/__data/assets/pdf_file/0004/162265/e96538.pdf

[76] Henderson KE, Recktenwald AJ, Reichley RM, et al. Clinical validation of the AHRQ postoperative venous thromboembolism patient safety indicator. Jt Comm J Qual Patient Saf. 2009;35(7):370-376. doi:10.1016/s1553-7250(09)35052-7

[77] https://github.com/MIT-LCP/mimic-code/tree/master/concepts

[78] Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care. 1998; 36: 8-27. doi: 10.1097/00005650-199801000-00004.

[79] Saber github repository, online, accessed 15 June 2020, https://github.com/BaderLab/saber/

[80] Xuezhe M, Hovy E. "End-to-end sequence labeling via bi-directional lstm-cnns-crf." arXiv preprint arXiv:1603.01354 (2016).

[81] Lample, Guillaume, et al. "Neural architectures for named entity recognition." arXiv preprint arXiv:1603.01360 (2016).

[82] Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40(Database issue): D940-6. doi: 10.1093/nar/gkr972.

[83] Crichton, G., Pyysalo, S., Chiu, B. et al. A neural network multi-task learning approach to biomedical named entity recognition. BMC Bioinformatics 2017;18: 368. https://doi.org/10.1186/s12859-017-1776-8

[84] Rehurek R, Sojka P. "Software framework for topic modelling with large corpora." In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010.

[85] Sievert C, Kenneth S. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014: 63-70. doi: 10.3115/v1/W14-3110.

[86] Engbers MJ, van Hylckama Vlieg A, Rosendaal FR. Venous thrombosis in the elderly: incidence, risk factors and risk groups. J Thromb Haemost. 2010; 8: 2105-2112. doi:10.1111/j.1538-7836.2010.03986.x

[87] Fernández-Delgado M et al. "Do we need hundreds of classifiers to solve real world classification problems?." The journal of machine learning research 2014; 15.1: 3133-3181.

[88] Tsamardinos Ioannis, Elissavet Greasidou, and Giorgos Borboudakis. "Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation." Machine Learning 107.12 2018: 1895-1922.

[89] Tsamardinos I, Charonyktakis P, Lakiotaki K, Borboudakis G, Zenklusen JC, et al. Just Add Data: Automated Predictive Modeling and BioSignature Discovery.

bioRxiv 2020.05.04.075747; doi: https://doi.org/10.1101/2020.05.04.075747

[90] Chawla NV, Bowyer KW, Hal LOl, Kegelmeyer WP. “SMOTE: synthetic minority over-sampling technique,”J. Artif. Intell.Res., 2002; 16: 321–357.

[91] Lemaitre G, Nogueira F, and Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research. 2017; 18: 559–563.

[92] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[93] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[94] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[95] https://developers.google.com/machine-learning/crash-course/classification/accuracy

[96] https://en.wikipedia.org/wiki/F-score

[97] Fortin Y, Crispo JAG, Cohen D, McNair DS, Mattison DR, et al. External validation and comparison of two variants of the Elixhauser comorbidity measures for all-cause mortality. PLOS ONE. 2017; 12(3): e0174379. https://doi.org/10.1371/journal.pone.0174379

[98] Danilatou V, Antonakaki D, Tzagkarakis C, Kanterakis A, Katos V, Kostoulas T. Automated mortality prediction in critically-ill patients with thrombosis using Machine Learning. IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2020).

[99] Spyropoulos AC, Anderson FA Jr, FitzGerald G, Decousus H, Pini M, et al; IMPROVE Investigators. Predictive and associative models to identify hospitalized medical patients at risk for VTE. Chest. 2011; 140(3): 706-714. doi: 10.1378/chest.10-1944.

[100] Barbar S, Noventa F, Rossetto V, Ferrari A, Brandolin B, Perlati M, De Bon E, Tormene D, Pagnan A, Prandoni P. A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score. J Thromb Haemost. 2010;8(11):2450-7. doi: 10.1111/j.1538-7836.2010.04044.x. PMID: 20738765.

[101] Viarasilpa T, Panyavachiraporn N, Marashi SM; Van Harn M; Kowalski RG, Mayer SA. Prediction of Symptomatic Venous Thromboembolism in Critically Ill Patients: The ICU-Venous Thromboembolism Score*, Critical Care Medicine. 2020; 48: p e470-e479 doi: 10.1097/CCM.0000000000004306.

[103] Nafee T, Gibson CM, Travis R, et al. Machine learning to predict venous thrombosis in acutely ill medical patients. Res Pract Thromb Haemost. 2020;4(2):230-237. doi:10.1002/rth2.12292

[104] Huisman MV, Klok FA. Current challenges in diagnostic imaging of venous thromboembolism. Blood. 2015 19; 126: 2376-82. doi: 10.1182/blood-2015-05-640979. PMID: 26585807.

[105] White RH, Zhou H, Romano PS. Incidence of symptomatic venous thromboembolism after different elective or urgent surgical procedures. Thrombosis and haemostasis. 2003; 90: 446–55.

[106] Veldhoen RA, Howes D, Maslove DM. Is Mortality a Useful Primary End Point for Critical Care Trials? Chest. 2020 Jul;158(1):206-211. doi: 10.1016/j.chest.2019.11.019.

[107] Sadeghi R, Banerjee T, Romine W. Early hospital mortality prediction using vital signals. Smart Health. 2018; 9-10: 265–274, 2018, CHASE 2018 Special Issue.

[108] Cugno M, Depetri F, Gnocchi L, Porro F, Bucciarelli P. Validation of the Predictive Model of the European Society of Cardiology for Early Mortality in Acute Pulmonary Embolism. TH Open. 2018; 2(3): e265-e271. doi: 10.1055/s-0038-1669427.

[109] Holmgren G, Andersson P, Jakobsson A, and A. Frigyesi. Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions. Journal of Intensive Care 2019; 7:44.

[110] Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM. Predicting Intensive Care Unit Readmission with Machine Learning Using Electronic Health Record ata. Ann Am Thorac Soc. 2018; 15(7): 846-853. Doi: 10.1513/AnnalsATS.201710-787OC.

[111] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006; 34(5): 1297-1310. doi:10.1097/01.CCM.0000215112.84523.F0

[112] Moreno RP, Metnitz PGH, Almeida E, 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 Med. 2005; 31: 1345-1355

[113] Johnson, Alistair E. W. BS1; Kramer, Andrew A. PhD2; Clifford, Gari D. PhD1 A New Severity of Illness Scale Using a Subset of Acute Physiology and Chronic Health Evaluation Data Elements Shows Comparable Predictive Accuracy*, Critical Care Medicine: 2013; 41: 1711-1718 doi: 10.1097/CCM.0b013e31828a24fe.

[114] Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. The Lancet Digital Health 2020; 2: e179-e191.

[115] Afessa B, Gajic O, Keegan MT. Severity of illness and organ failure assessment in adult intensive care units. Crit Care Clin 2007; 23: 639-58.

[116] Komorowski M, Celi LA. Will Artificial Intelligence Contribute to Overuse in Healthcare? Crit Care Med. 2017;45(5):912-913. Doi:10.1097/CCM.0000000000002351.

[117] Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: Observational Study. JMIR Medical Informatics 2019; 7.

[118] Fernando SM, Mok G, Castellucci LA, Dowlatshahi D, Rochwerg B,et al. Impact of Anticoagulation on Mortality and Resource Utilization Among Critically Ill Patients With Major Bleeding. Crit Care Med. 2020; 48: 515-524. doi: 10.1097/CCM.0000000000004206.

[119] Fernandez R, Cano S, Catalan I, Rubio O, Subira C, et al. High red blood cell distribution width as a marker of hospital mortality after ICU discharge: a cohort study. J Intensive Care. 2018;6:74. doi: 10.1186/s40560-018-0343-3.

[120] Wong CCY, Chow WWK, Lau JK, Chow V, Ng ACC, Kritharides L. Red blood cell transfusion and outcomes in acute pulmonary embolism. Respirology. 2018; 23: 935-941. doi: 10.1111/resp.13314.

[121] Arihan O, Wernly B, Lichtenauer M, Franz M, Kabisch B et al. Blood Urea Nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU. PLoS One. 2018; 13(1): e0191697. doi:10.1371/journal.pone.0191697

[122] Salvagno GL, Sanchis-Gomar F, Picanza A, Lippi G. Red blood cell distribution width: A simple parameter with multiple clinical applications. Crit Rev Clin Lab Sci. 2015; 52(2): 86-105. doi: 10.3109/10408363.2014.992064.

[123] Santos MS, Soares JP, Abreu PH, Araujo, H, Santos J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. IEEE Comp. Intell. Mag. 2018; 13: 59–76.

[124] Goel R, Patel EU, Cushing MM, Frank SM, Ness PM, Takemoto CM, Vasovic LV, Sheth S, Nellis ME, Shaz B, Tobian AAR. Association of Perioperative Red Blood Cell Transfusions With Venous Thromboembolism in a North American Registry. JAMA Surg. 2018; 153(9): 826-833.

[127] Zheng Y, Lu C, Wei S, Li Y, Long L, Yin P. Association of red blood cell transfusion and in-hospital mortality in patients admitted to the intensive care unit: a systematic review and meta-analysis. Crit Care. 2014; 18(6): 515. doi:10.1186/s13054-014-0515-z

[128] Divatia JV, Khan PU, Myatra SN. Tracheal intubation in the ICU: Life saving or life threatening?. Indian J Anaesth. 2011; 55(5): 470-475. doi:10.4103/0019-5049.89872

[129] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B. Warde-Farley D et al. Generative Adversarial Networks. Proceedings of the International Conference on Neural Information Processing Systems. 2014: 2672–2680.

[130] Moor M, Rieck B, Horn M, Jutzeler C, Borgwardt K. Early Prediction of Sepsis in the ICU using Machine Learning: A Systematic Review.medRxiv 2020.08.31.20185207; doi:https://doi.org/10.1101/2020.08.31.20185207.

[131] Keuning BE, Kaufmann T, Wiersema R, Granholm A, Pettilä V, Møller MH, Christiansen CF, Castela Forte J, Snieder H, Keus F, Pleijhuis RG, van der Horst ICC; HEALICS consortium. Mortality prediction models in the adult critically ill: A scoping review. Acta Anaesthesiol Scand. 2020;64(4):424-442. doi: 10.1111/aas.13527.

[132] Min H, Avramovic S, Wojtusiak J, Khosla R, Fletcher RD, Alemi F, Kheirbek RE. A Comprehensive Multimorbidity Index for Predicting Mortality in Intensive Care Unit Patients. J Palliat Med. 2017; 20: 35-41. doi: 10.1089/jpm.2015.0392.

[133] Sabra S, Mahmood Malik K, Alobaidi M. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. Comput Biol Med. 2018; 94: 1-10. doi:10.1016/j.compbiomed.2017.12.026.

[134] Bhavani Singh, A. K., et al. "Multi-label natural language processing to identify diagnosis and procedure codes from MIMIC-III inpatient notes." arXiv (2020): arXiv-2003.

[135] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018 Sep 11; 5: 180178. doi: 10.1038/sdata.2018.178.

[136] Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform. 2018; 83: 112-134. doi:10.1016/j.jbi.2018.04.007.