Automated Mortality Prediction in Critically-ill Patients with Thrombosis using Machine Learning

V. Danilatou*†, D. Antonakaki‡, C. Tzagkarakis‡, A. Kanterakis‡, V. Katos*, T. Kostoulas*
*Bournemouth University, Faculty of Science and Technology, Bournemouth, UK
†Venizeleio Hospital of Heraklion, Heraklion, Greece
‡Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH)
vdanilatou@bournemouth.ac.uk, {kantale,tzagarak,despoina}@ics.forth.gr, {vkatos,tkostoulas}@bournemouth.ac.uk

***Abstract*—Venous thromboembolism (VTE) is the third most common cardiovascular condition. Some high risk patients diagnosed with VTE need immediate treatment and monitoring in intensive care units (ICU) as the mortality rate is high. Most of the published predictive models for ICU mortality give information on in-hospital mortality using data recorded in the first day of ICU admission. The purpose of the current study is to predict in-hospital and after-discharge mortality in patients with VTE admitted to ICU using a machine learning (ML) framework.**

We studied 2,468 patients from the Medical Information Mart for Intensive Care (MIMIC-III) database, admitted to ICU with a diagnosis of VTE. We formed ML classification tasks for early and late mortality prediction. In total, 1,471 features were extracted for each patient, grouped in seven categories each representing a different type of medical assessment. We used an automated ML platform, JADBIO, as well as a class balancing combined with a Random Forest classifier, in order to evaluate the importance of class imbalance. Both methods showed significant ability in prediction of early mortality (AUC=0.92). Nevertheless, the task of predicting late mortality was less efficient (AUC=0.82).

To the best of our knowledge, this is the first study in which ML is used to predict short-term and long-term mortality for ICU patients with VTE based on a multitude of clinical features collected over time.

***Index Terms*—MIMIC-III, ICU mortality prediction, thrombosis, machine learning, imbalanced classification, AutoML**

I. Introduction

Venous thromboembolism (VTE) that presents with clots in the veins, most frequently as deep vein thrombosis (DVT) or pulmonary embolism (PE) is a potentially lethal disease with an annual prevalence rate of approximately 1 per 1000 adults [1]. Its prevalence is even higher in hospitalized, critically-ill and cancer patients [2]. In critically ill patients, it is associated with significant morbidity, prolonged intensive care units (ICU) and hospital stay and increased in-hospital and post-discharge morbidity and mortality [3].

VTE is a complex multifactorial disease. Besides hereditary, common strong acquired risk factors are surgery, congestive heart or respiratory failure, cancer, and trauma. Hospitalization, increasing age, and obesity are also considered but as weak risk factors [4].

Several prognostic models that incorporate clinical and 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 (sPESI) for pulmonary embolism [5]. Moreover, there are several other scores, such as Sequential Organ Failure Assessment (SOFA) [6], Oxford Acute Severity of Illness Score (OASIS) [7], Acute Physiology And Chronic Health Evaluation (APACHE) [8], Simplified Acute Physiology Score (SAPS) [9], that estimate the severity of disease in ICU and correlate positively mostly with early mortality but with varying accuracy depending on the population studied. These scores are based on data obtained during the first day of admission so they lack considerable information stemming during their hospital stay and post-discharge. Moreover, they are not widely customised in different patient groups, such as patients with thrombosis. 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 prompt recognition or adequate treatment could probably improve survival [10]. With the recent advancements in electronic health records, big data storage, and machine learning (ML) algorithms it is possible to build forecasting systems to guide clinicians making more informed predictions [11].

MIMIC-III database is a freely accessible database that provides detailed granular clinical data and gives the opportunity for data sharing, code sharing and ML benchmarking [12]. Several studies have been published regarding mortality prediction using ML but they are mostly based on the extraction of a simple feature set and patients admitted in the ICU regardless of the primary diagnosis [13]. Since diagnosis could significantly affect survival it would be interesting to study disease-related outcomes and try to identify specific clinical features with prognostic significance. Even more importantly, it is necessary to predict post-discharge mortality which is a difficult task, since patients admitted to ICU usually suffer from a high comorbidity burden. Few studies have focused on the prediction of late mortality using ML algorithms [10].

Here we present a thorough ML analysis pipeline with the purpose of predicting early as well as late mortality after ICU admission for VTE. The novelty of our approach is that we used a very wide selection of features from the MIMIC-III database. This includes demographic information, prescriptions, procedures, comorbidity and severity scores as well as information coming from written notes. This imposes dealing with a variety of pre-processing techniques, in order to account for the different types of features. Having in our disposal a feature rich dataset, allows us to examine

how current comorbidity and severity scores are associated between each other, and if they have high predictive ability for mortality when compared with common clinical features. In order to tackle the "curse of dimensionality" issue in a dataset of this shape we apply an "AutoML" technique [14] that is tailored for low-sample, high dimensional, imbalanced and sparse (many missing values) data. Finally, we examined the class balancing effect in light of oversampling combined with a Random Forest classifier in both prediction tasks, since JADBIO addresses imbalanced classes through stratified cross validation and diversified class weights during Support Vector Machine (SVM) learning [14].

II. Materials and Methods

A. Data Source

Data were obtained from Medical Information Mart for Intensive Care database (MIMIC-III, version 1.4) that comprises 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. [12]. Diagnosis is given as primary and secondary diagnosis ICD-9 codes as well as diagnosis-related groups (DRG).

B. Ethics Statement

MIMIC-III database was created in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards and all investigators with data access (VD, CT) were approved by PhysioNet. Patient data were de-identified and date-shifted. All pre-processing and data analysis was performed under MIMIC-III regulations.

C. Dataset Description

Selection of patients was based on 35 different ICD-9 codes related to thrombosis. Validation of this grouping for thrombosis diagnosis from an independent panel of physicians showed very good performance [15]. Overall 2,468 patients were selected for our study (6.4% of total patients in MIMIC-III). Patients younger than 15 years (n=3), pregnancy and puerperium complications (n=40) and patients with "do not resuscitate code" (DNR) (n=169) were excluded. Three main groups of thrombotic diseases were recognised: pulmonary embolism (n=960), deep vein thrombosis and thrombophlebitis (n=1,543) and unusual site thrombosis (n=307). Many patients belonged in more than one diagnostic category, as shown in Fig. 1. All VTE patients were split into three groups. The first, referred as G_1 are 348 patients that died during the first ICU admission in which they were diagnosed with thrombosis. Patients in this group died on average 17 days after their admission with a median of 11 days. The second, referred as G_2 are 817 patients that died after their discharge from ICU or in a later admission. On average this group died after 549 days with a median of 225 days. The third, referred as G_3 are 1,303 patients that remained alive for months after their admission in ICU. From these groups we form two ML tasks. The first is to build a model that distinguishes G_1 vs. G_3 patients (called "early mortality" or M_1) and the second is a model that distinguishes G_2 vs. G_3 patients (called "late mortality" or M_2).

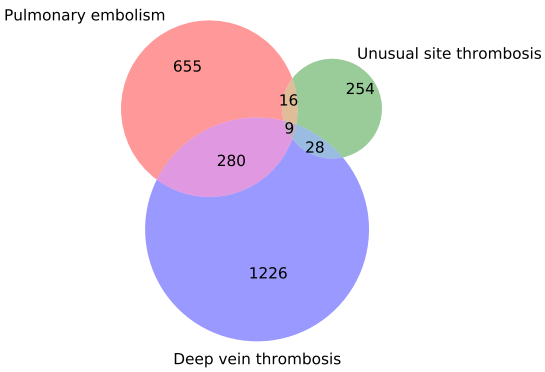

Fig. 1. Venn diagram showing the number of patients belonging to all subsets of the three diagnostic categories.

The clinical characteristics of the study population are described in Table I. From the characteristics presented in this Table, 9 were added as features.

D. Feature Selection

For each of these patients we extracted a multitude of features based on factors that could be associated with thrombosis. Since our purpose was to investigate potential novel discriminatory features, we chose to be very liberal on feature extraction from the database. 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 features were grouped in seven categories each representing a different type of medical assessment or interventions as shown in Table II. In cases where features had time-series data the first measurement and average were extracted. Concepts are meta-features containing the values of various scores and measurements. These values are not stored in the database but are available as SQL queries that estimate them from other features. Concepts include a set of severity illness and organ failure scores such as SOFA, SAPS, Glasgow Coma Scale (GCS), sepsis scores (Martin, Angus). and comorbidity scores that are described as different Elixhauser indices [16].

NoteEvents contains unstructured notes written by clinicians in free text format. These notes have been proved to contain valuable information that when combined with semantic and sentiment analysis can be even used for predicting VTE [17]. Here, our objective was to convert this textual information in numerical that could be added in our feature set. Towards this direction, first, we extracted all clinically relevant entities from the text using the SABER sequence annotator[1] which is a Deep Neural Network framework, tailored for entity extraction from

[1]https://github.com/BaderLab/saber/

TABLE I
DEMOGRAPHIC AND CLINICAL CHARACTERISTICS OF 2,468 ICU PATIENTS WITH THROMBOSIS IN MIMIC-III DATABASE. (+): THIS CHARACTERISTIC WAS ADDED AS A FEATURE.

Characteristic	Value	Characteristic	Value
Overall patients with thrombosis	2,468	Length of ICU stay (LOS), days (+)	
• Pulmonary embolism (+)	960 (38.9%)	• Average (SD)	7.06 (10.06)
• Deep vein thrombosis (+)	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 (+)		Cancer diagnosis (+)	605 (24.5%)
• White	1,801 (73%)	Mortality (%)	
• Black	246 (10%)	• G_1 or Early (at the first admission)	348 (14.1%)
• Other	421 (17%)	• G_2 or Late (1 year mortality)	817 (33.1%)
		• G_3 or "Alive"	1303 (52.8%)
Age, years (+)		Time to death (in days)	
• Average (SD)	62.64 (16.7) [min=17.4 max=98.7]	• Average (SD)	390 (647)
		• Median	83

biomedical documents. SABER uses a bi-directional Long Short-Term Memory (LSTM) architecture [18] and provides access to pre-trained models for various types of entities. One of these is the disease ontology (DO) [19] which is a structured vocabulary of entities related to various pathologies and symptoms. For each NoteEvent entry we extracted all DO entities. On average for each patient we extracted 161 entities with a median of 133. Next, we fitted these entities into a Latent Dirichlet Allocation (LDA) topic model with the Gensim framework [20] by using 50 topics.

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. As a result, this produced a 50 dimensional space that contained the topic distribution for each patient. An example of the visualization of this model with the LDAvis tool [21] can be found in the following URL: https://doi.org/10.6084/m9.figshare.12854852.v1. Thus, 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.

In total our dataset contained 1,471 features: 9 clinical (Table I) and 1,462 features from the groups presented here (Table II). It is obvious that each group describes a different view of the clinical picture of the patient. Since our objective is to locate subsets of discriminatory features, we applied a stratified analysis for each group. Thus, for each ML task, we created subsets 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. We also created datasets that contained the entirety of the features. In total we created 16 datasets which correspond to the two ML tasks combined with the 8 groupings (seven groups plus one containing all groups).

E. AutoML

For each of these 16 datasets we applied a classification ML pipeline. For this purpose, we used the JADBIO AutoML platform that uses an Artificial Intelligence (AI) Decision Support System called Algorithm and Hyper-Parameter Space selection (AHPS) in order to extract predictive models and signatures. JADBIO works as follows: initially it constructs a set of ML configurations consisting of algorithms and hyper-parameters. The algorithms are Linear, Ridge and Lasso Regression (LR), Decision Trees (DT), 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 [22]. Subsequently, it evaluates these configurations through a bootstrap corrected cross-validation algorithm [23]. After selecting the "winning configuration" it reports the classification statistics like truth table, AUC, sensitivity, specificity, precision, selected features along with their classification ability and their 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 [14]. 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 [14].

F. Class Balancing Based on Oversampling

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 (here we adopt the Random Forest (RF) classifier) both for M_1 and M_2 tasks. More specific, the imbalance ratio for M_1 is 1:3.744, and 1:1.594 for M_2. We adopt

TABLE II
DESCRIPTION OF CLINICAL AND LABORATORY FEATURES SELECTED FROM MIMIC-III DATABASE. THE FIRST COLUMN DESCRIBES THE CORRESPONDING TABLE FROM THE MIMIC-III DATABASE. ABBREVIATIONS: RBC=RED BLOOD CELL, PLT=PLATELET.

Group	Description	#features	Average	Median	Most common features
ChartEvents	Vital signs, labs, clinical information	235	433	77	Common labs, blood gases, blood pressure
LabEvents	Laboratory indices	45	1237	1157	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
InputeEvents	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
NoteEvents	Unstructured medical notes	50	48 entries, 2408 characters	1382 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, glasgow coma scale, first day laboratories, first day vital signs, transfusions

the Synthetic Minority Oversampling Technique (SMOTE) method [24], where we use the default SMOTE implementation `sm=SMOTE(random_state=random_seed)` included in the Imbalanced-Learn [25] Python package[2]. Prior to class balancing, we follow the next steps. First, we drop the features (columns) from both datasets that have more than 50% percent of missing values. Second, the boolean values are replaced as `TRUE`: 1, `FALSE`: 0, and the gender (male/female) as well as the ethnicity (white/black/other) feature is one-hot encoded. Third, median imputation is adopted to fill the missing values. A shuffled stratified 75%/25% train/test split is applied on M_1 or M_2 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 our focus is given on examining the SMOTE oversampling effect on the final performance evaluation, we apply the SMOTE 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 [26]. 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 overoptimism. As a result, four distinct cases arise: 5-fold stratified cross-validation on M_1 or M_2 training partition with or without SMOTE oversampling. In both cases, we perform grid-search hyper-parameter tuning of a RF classifier which is robust and efficient when dealing with numerical, categorical and boolean data.

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 hyper-parameters combination. Then, we train the best (F1-based selected) RF model on the entire initial (before the cross-validation iterations) training partition. Towards the final performance evaluation, we compute the average ROC curves, 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.

[2]https://imbalanced-learn.readthedocs.io/en/stable/index.html

III. RESULTS

A. Correlation of Sepsis, Comorbidities and Organ Failure Scores

First, we analyzed the complex interactions between the various sepsis (n=20), comorbidities (n=17) and organ failure (n=12) scores. This comparison tries to examine the level on which various scores are complementary and to which extent correlate with each other [27]. For each of these score groups we computed a Pearson pairwise correlation matrix and we visualized these correlations using heatmaps. For sepsis and severity scores we added an extra feature which was the "time before death" containing the negative of the time (in days) in which the patient died after their first admission with a thrombosis diagnosis. Regarding sepsis (see Fig.2), it is interesting that there is a quite good correlation between the two sepsis scores (Angus and Martin). As far as it concerns comorbidity index (data not shown) we used the Quan Elixhauer score, since both variants of Elixhauer measures AHRQ and Quan have comparable efficiency in predicting all-cause mortality [28]. Finally, we observe a strong correlation between various severity and organ failure scores as shown in Fig.3, although only SAPS and Acute Physiology Score (APS) might have a weak correlation with time to death.

B. Classification of Early and Late Mortality Patients

The best ML model chosen by JADBIO to predict early mortality (task M_1) was RF training 500 trees with Deviance splitting criterion and minimum leaf size equal to 2. As expected, the best performance corresponded to the

dataset containing all groups (AUC=0.925), followed by Concepts (AUC=0.923) and ChartEvents (AUC=0.917), whereas InputEvents had the worst performance (AUC=0.781) (see Fig.4). Regarding late mortality (task M_2), the best ML model was again RF training 500 trees with Deviance splitting criterion and minimum leaf size equal to 3. Nevertheless, the task of predicting M_2 was less efficient even with the holistic approach (AUC=0.82).

As it is shown in Fig.5, Concepts in this case had inferior performance (AUC=0.783) which is expected since known severity and organ failure scores are excellent only for predicting M_1. This difference can also be attributed to the fact that an unknown number of patients in the "alive" (G_3) group might in fact have the same mortality risk as in the patients in G_2 group due to the limited time period that the database tracks mortality status. Another interesting finding is that NoteEvents (free text features) had almost the same AUC (0.762) as ChartEvents (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.

C. Mortality Prediction Based on SMOTE and Random Forests

Fig. 6 depicts the average ROC curves in the case of M_1 (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. Since the imbalance ratio in the case of M_2 is even lower (i.e., 1:1.595) we expect 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 Fig. 6 (dashed lines) where the mean ROC scores are 0.81 and 0.82 in the case of SMOTE and no-SMOTE, respectively. In general oversampling techniques such as SMOTE perform better in high imbalance ratio

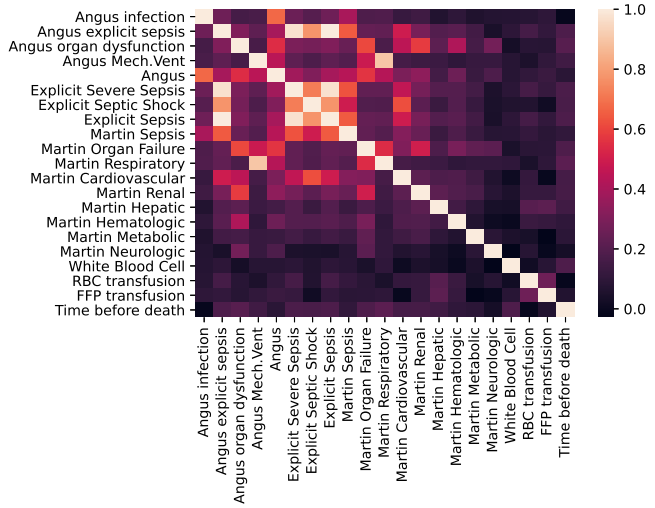

Fig. 2. Feature correlation results for sepsis.

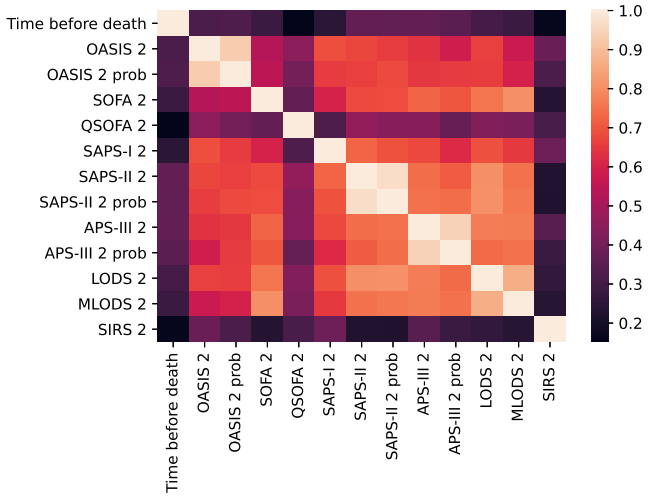

Fig. 3. Feature correlation results for severity scores.

datasets [26], [29]. Therefore, we can conclude that there is no drastic change in the early and late mortality predictive accuracy regardless of a class balance oversampling use within a ML pipeline.

D. Feature Discriminative Analysis

Cancer and age at thrombosis were significant predictors in most of the analysis subgroups for early as well as late mortality. Anticoagulation with warfarin in "All" and "Prescriptions" was another significant predictor for both M_1 and M_2. Selected features to predict M_1 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

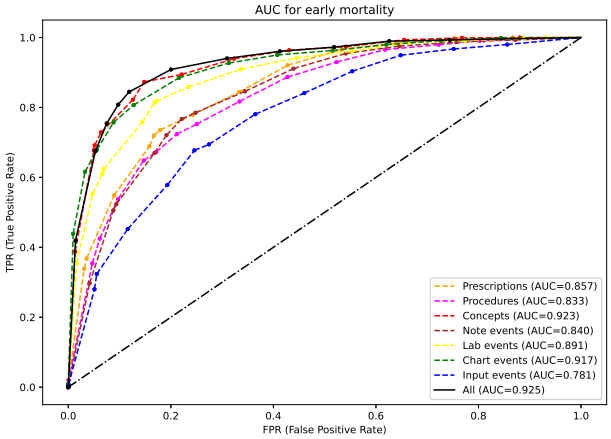

Fig. 4. AUC for early mortality based on JADBIO.

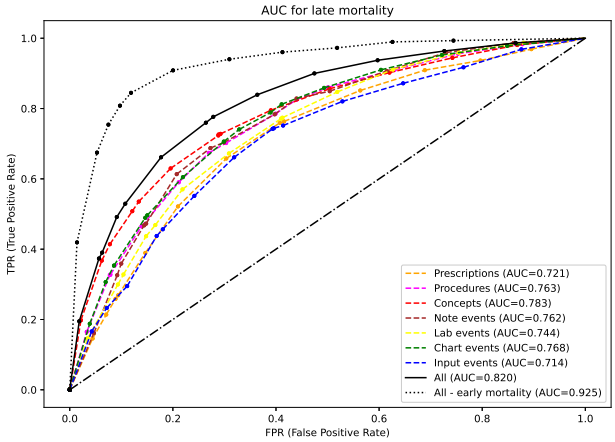

Fig. 5. AUC for late mortality based on JADBIO.

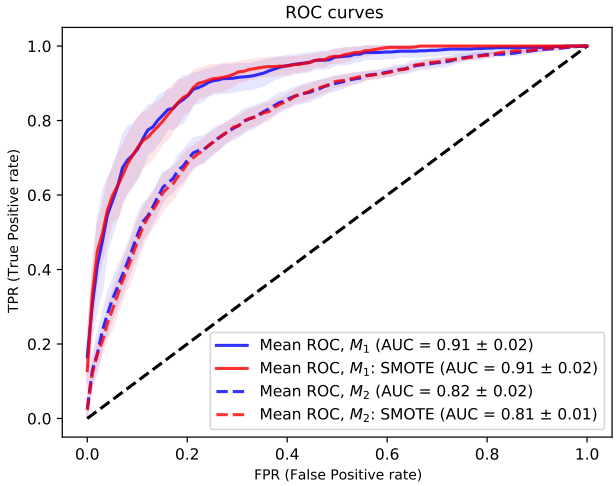

Fig. 6. AUC for early and late mortality based on SMOTE and RF.

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-RDW, platelets, white blood cells) were recognised in "LabEvents" and "All" datasets, as shown in Fig. 7. Selected predictive features for late mortality were similarly associated with cardiovascular and renal failure, medications and laboratory indices. Renal failure was indicated by creatinine avg, 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 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".

IV. Discussion

Prediction of early and late mortality in ICU patients has been a central challenge in the area of medical informatics. Current approaches use either a limited pre-selected number of features [13], [30], [31] or explore the feature space with a small range of ML algorithms (i.e., Logistic Regression [32], SVM [33], Artificial Neural Networks [34], Decision Trees [35]). Even when more generic approaches are used it is questionable whether proper ML guidelines for overfitting prevention and accurate efficiency metrics reporting are followed. Here, we focus on predicting early and late mortality for patients admitted in ICU with VTE diagnosis. The main goal of this work is to locate features and build models in order to improve ICU survival rates.

Features included in our study derived from raw clinical measurements, widely accepted severity, comorbidity and organ failure scores (presented here as Concepts) [6]–[11], as well as information coming from free-text notes. This 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 [36] and co-dependencies between different features.

To tackle these issues we employed two strategies. The first is an AutoML approach based on JADBIO, that has been widely tested in biomedical data and follows all good practices for analysis and efficiency reporting. Besides, JADBIO can produce "interpretable" models that can be intuitively explored and explained by physicians as confirmed by our study. 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 [30]. Similarly with [30] we did not find sex to be a predictor of ICU mortality. Moreover, individual feature analysis confirmed that warfarin [37], RDW [38] and red blood cell transfusions [39] are significant predictors of early and probably of long-term mortality. Our second approach is a class balancing combined with a RF classifier approach, indicating that the results we obtain from JADBIO are consistent in terms of class imbalance.

Our results show that 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 as in our case, from 0.783 to 0.82. 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 [32]. Our model outperforms [31] in prediction of early mortality (AUC 0.92 vs 0.77) Also in a recent review [40], 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 [32] had an AUC of 0.86 for 6 month mortality and 0.88 for 12 month mortality. Regarding ICU scores, [10] reports AUC of 0.826, 0.836, and 0.788 for SAPS II, APACHE II, and SOFA scales,

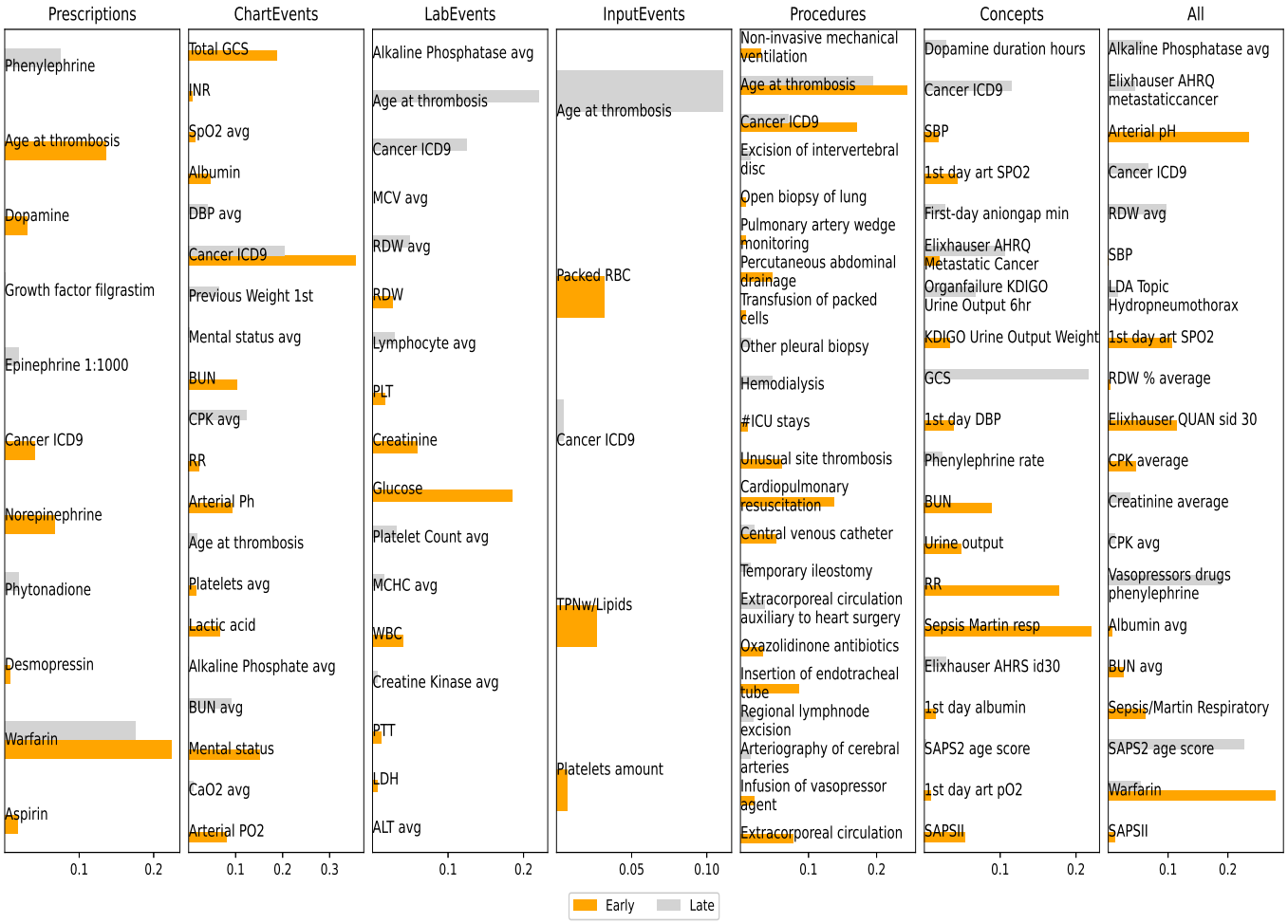

Fig. 7. The discriminative power for the top features selected from JADBIO for each group except NoteEvents. Values represent the relative change of the AUC. Orange bars represent features for Early mortality and Gray bars represent features for Late mortality. ALT=Alanine Aminotransferase, Art pO2=Arterial Oxygen Partial Pressure, 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 Latent Dirichlet Allocation, 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

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. Our correlation study confirmed the hypothesis that different sepsis and comorbidity scores convey different types of information [27].

Some limitations of our study should be taken into account. Our study was monocentric and retrospective. Since the data were collected in the past, it is possible that many medical practices have changed over time (e.g. warfarin). Selection of our population was based solely on ICD-9 codes and DRG codes. This could include some false negative and false positive cases since confirmation by imaging studies was not feasible. More importantly, no external validation of our results has been performed. Finally, we believe that a more focused approach of semantic extraction could be more effective [17].

Our future work includes external validation of our model in eICU Collaborative Research Database, which is a larger and more recent database [41]. Regarding extraction of specific ontology based entities, another alternative would be to use direct language embeddings [42]. Finally we plan to investigate the use of LSTM for importing time-series data in our model [43].

In conclusion, early mortality in critically-ill patients with VTE can be easily predicted by automated ML. There is a need for more precise and reliable tools in order to estimate late mortality in VTE patients successfully discharged from the ICU.

Acknowledgments

We would like to thank Prof. Ioannis Tsamardinos, Vincenzo Lagani and Naomi Thomson for their help and support on using JADBIO. This work has been partially supported by IDEAL-CITIES; a European Union's Horizon 2020 research and innovation staff exchange programme (RISE) under the Marie Skłodowska-Curie grant agreement No 778229.

REFERENCES

[1] J. A. Heit, "Epidemiology of venous thromboembolism," *Nature Reviews Cardiology*, vol. 12(8), pp. 464–474, 2015.

[2] M. Bahloul *et al.*, "Pulmonary embolism in intensive care unit: predictive factors, clinical manifestations and outcome," *Annals of Thoracic Medicine*, vol. 5, pp. 97–103, 2010.

[3] D. R. Hirsch, E. P. Ingenito, and S. Z. Goldhaber, "Prevalence of deep venous thrombosis among patients in medical intensive care," *JAMA*, vol. 274, pp. 335–337, 1995.

[4] F. Moheimani and D. Jackson, "Venous thromboembolism: classification, risk factors, diagnosis, and management," *ISRN Hematology*, vol. 2011, no. 124610, 2011.

[5] D. Jiménez, D. Aujesky, L. Moores *et al.*, "Simplification of the pulmonary embolism severity index for prognostication in patients with acute symptomatic pulmonary embolism," *Arch Intern Med.*, vol. 170, no. 15, pp. 1383–1389, 2010.

[6] F. L. Ferreira, D. P. Bota, A. Bross, C. Mélot, and J. L. Vincent, "Serial evaluation of the SOFA score to predict outcome in critically ill patients," *JAMA*, vol. 286, no. 14, pp. 1754–1758, 2001.

[7] A. E. W. Johnson, A. A. Kramer, and G. D. Clifford, "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.*, vol. 41, no. 7, pp. 1711–1718, 2013.

[8] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients," *Crit. Care Med.*, vol. 34, no. 5, pp. 1297–1310, 2006.

[9] J. L. Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *JAMA*, vol. 270, no. 24, pp. 2957–2963, 1993.

[10] P. A. Fuchs, I. J. Czech, and Ł. J. Krzych, "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*, vol. 16, no. 18, 2019.

[11] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *The New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[12] A. Johnson, T. Pollard, L. Shen *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, 2016.

[13] R. Sadeghi, T. Banerjee, and W. Romine, "Early hospital mortality prediction using vital signals," *Smart Health*, vol. 9-10, pp. 265–274, 2018, cHASE 2018 Special Issue.

[14] I. Tsamardinos, P. Charonyktakis, K. Lakiotaki, G. Borboudakis, J. C. Zenklusen, H. Juhl, E. Chatzaki, and V. Lagani, "Just add data: Automated predictive modeling and biosignature discovery," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/05/05/2020.05.04.075747

[15] K. E. Henderson, A. J. Recktenwald, R. M. Reichley *et al.*, "Clinical validation of the AHRQ postoperative venous thromboembolism patient safety indicator," *The Joint Commission Journal on Quality and Patient Safety*, vol. 35, no. 7, pp. 370–376, 2009.

[16] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical Care*, vol. 36, no. 1, pp. 8–27, 1998.

[17] S. Sabra, K. M. Malik, and M. Alobaidi, "Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives," *Computers in Biology and Medicine*, vol. 94, pp. 1–10, 2018.

[18] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. of the 54th Annual Meeting of the Assoc. for Comp. Ling.*

[19] L. M. Schriml *et al.*, "Disease ontology: a backbone for disease semantic integration," *Nucleic Acids Res.*, vol. 40, pp. D940—-D946, 2012.

[20] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, 2010, pp. 45–50.

[21] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proc. of the Workshop on Interactive Language Learning, Visualization, and Interfaces*.

[22] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[23] I. Tsamardinos, E. Greasidou, and G. Borboudakis, "Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation," *Machine Learning*, vol. 107, no. 12, pp. 1895–1922, 2018.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[25] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[26] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]," *IEEE Comp. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, 2018.

[27] J. Vincent and R. Moreno, "Clinical review: Scoring systems in the critically ill," *Crit Care.*, vol. 14, no. 2, 2010.

[28] Y. Fortin, J. A. G. Crispo *et al.*, "External validation and comparison of two variants of the Elixhauser comorbidity measures for all-cause mortality," *PLoS One*, vol. 12, no. 3, 2017.

[29] J. Kong, W. Kowalczyk, D. A. Nguyen, T. Bäck, and S. Menzel, "Hyperparameter optimisation for improving classification under class imbalance," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 3072–3078.

[30] K. M. Ho, M. Knuiman, J. Finn, and S. A. Webb, "Estimating long-term survival of critically ill patients: The PREDICT model," *PLoS ONE*, vol. 3, no. 9, 2008.

[31] M. Cugno, F. Depetri, L. Gnocchi, F. Porro, and P. Bucciarelli, "Validation of the predictive model of the european society of cardiology for early mortality in acute pulmonary embolism," *TH Open*, vol. 2, no. 3, pp. 265–271, 2018.

[32] H. Min, S. Avramovic, J. Wojtusiak, R. Khosla, R. D. Fletcher, F. Alemi, and R. Elfadel Kheirbek, "A comprehensive multimorbidity index for predicting mortality in intensive care unit patients," *Journal of palliative medicine*, vol. 20, no. 1, pp. 35–41, 2017.

[33] P. Ferroni, F. M. Zanzotto, N. Scarpato *et al.*, "Risk assessment for venous thromboembolism in chemotherapy-treated ambulatory cancer patients: A machine learning approach," *Med. Dec. Making*, vol. 37, no. 2, pp. 234–242, 2017.

[34] G. Holmgren, P. Andersson, A. Jakobsson, 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*, vol. 7:44, 2019.

[35] J. C.Rojas, K. A. Carey, D. P. Edelson, L. R. Venable, M. D. Howell, and M. M. Churpek, "Predicting intensive care unit readmission with machine learning using electronic health record data," *Annals of the Am. Thorac. Soc.*, vol. 15, no. 7, pp. 846–853, 2018.

[36] A. Sharafoddini, J. A. Dubin, D. M. Maslove, and J. Lee, "A new insight into missing data in intensive care unit patient profiles: Observational study," *JMIR Medical Informatics*, vol. 7, no. 1, 2019.

[37] S. M. Fernando, G. Mok, L. A. Castellucci *et al.*, "Impact of anticoagulation on mortality and resource utilization among critically ill patients with major bleeding," *Crit. Care Med.*, vol. 48, no. 4, pp. 515–524, 2020.

[38] R. Fernandez, S. Cano, I. Catalan *et al.*, "High red blood cell distribution width as a marker of hospital mortality after ICU discharge: a cohort study," *Journal of Intensive Care*, vol. 6, no. 74, 2018.

[39] C. C. Y. Wong, W. W. K. Chow, J. K. Lau, V. Chow, A. C. C. Ng, and L. Kritharides, "Red blood cell transfusion and outcomes in acute pulmonary embolism," *Journal of the Asian Pac. Soc. of Respirology*, vol. 23, no. 10, pp. 935–941, 2018.

[40] B. E. Keuning, T. Kaufmann, R. Wiersema *et al.*, "Mortality prediction models in the adult critically ill: A scoping review," *Acta Anaesthesiol. Scand.*, vol. 64, no. 4, pp. 424–442, 2020.

[41] T. J. Pollard *et al.*, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5:180178, 2018.

[42] A. K. B. Singh *et al.*, "Multi-label natural language processing to identify diagnosis and procedure codes from MIMIC-III inpatient notes," *arXiv:2003.07507*, 2020.

[43] H.-C. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen *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," *Lancet Digital Health*, vol. 2, pp. 179–191, 2020.