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

V. Danilatou*, D. Antonakaki†, C. Tzagarakis‡, A. Kanterakis‡, V. Katos*, T. Kostoulas*
*Bournemouth University, Faculty of Science and Technology, Bournemouth, UK
†Venizelio 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,dkostoulas}@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 JADDBIO 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 $M_2$).

The clinical characteristics of the study population are described in Table [1]. 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 [1]. 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 [18] which is a Deep Neural Network framework, tailored for entity extraction from

https://github.com/BaderLab/saber/
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 I). 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 $M_2$ is 1:3.744, and 1:1.594 for $M_2$. We adopt
In both cases, we perform grid-search hyper-parameter tuning with different realizations of the train/test split, the 5-fold stratified cross-validation, and randomizations of the SMOTE method.

### 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. During cross-validation, we compute the average ROC curves, where the results are averaged over ten Monte Carlo repetitions (cross-validation iterations) training partition. Towards the final performance evaluation, we compute the average ROC curves, with or without SMOTE oversampling.

The best ML model chosen by JADBIO to predict early mortality (task $M_1$) was RF training 500 trees with De-

### 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 De-

---


---

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>#features</th>
<th>Average</th>
<th>Median</th>
<th>Most common features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChartEvents</td>
<td>Vital signs, labs, clinical information</td>
<td>235</td>
<td>433</td>
<td>77</td>
<td>Common labs, blood gases, blood pressure</td>
</tr>
<tr>
<td>LabEvents</td>
<td>Laboratory indices</td>
<td>45</td>
<td>1237</td>
<td>1157</td>
<td>Hematocrit, hemoglobin, white blood cells, platelets, red blood cells, renal and liver function tests, hemostasis screening tests, sepsis indices</td>
</tr>
<tr>
<td>Procedures</td>
<td>Several procedures including transfection and mechanical ventilation</td>
<td>526</td>
<td>24.3</td>
<td>6</td>
<td>Venous catheterization, enteral nutrition, endotracheal intubation, mechanical ventilation for more than 96 hours</td>
</tr>
<tr>
<td>InputeEvents</td>
<td>Transfusion and parenteral nutrition</td>
<td>12(MV)</td>
<td>10(CV)</td>
<td>RBC transfusion, PLT transfusions, plasma transfusions</td>
<td></td>
</tr>
<tr>
<td>Prescriptions</td>
<td>Medications</td>
<td>91</td>
<td>132</td>
<td>14</td>
<td>Heparin, insulin, warfarin, aspirin, enoxaparin, norepinephrine, phytonadione and atorvastatin</td>
</tr>
<tr>
<td>NoteEvents</td>
<td>Unstructured medical notes</td>
<td>50</td>
<td>48 entries, 2408 characters</td>
<td>N/A</td>
<td>Comorbidity indices, severity illness scores, organ failure scores, sepsis scores, Glasgow coma scale, first day laboratories, first day vital signs, transfusions</td>
</tr>
<tr>
<td>Concepts</td>
<td>Scores, first day labs, first day vital signs, doses and durations of medications</td>
<td>493</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
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.3). Regarding late mortality (task M2), 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 M2 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 M1. This difference can also be attributed to the fact that an unknown number of patients in the “alive” (G1) 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. 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 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. Since the imbalance ratio in the case of M2 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. 4 (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 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 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
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.
Some limitations of our study should be taken into account. Our future work includes external validation of our model in a more recent database [41]. Regarding extraction of specific concepts, 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 Sklodowska-Curie grant agreement No 778229.