Sampling Methods and Feature Selection for Mortality Prediction with Neural Networks

Christian Steinmeyer*, Dr. Lena Wiese† *Research Group Bioinformatics*
Fraunhofer Institute for Toxicology and Experimental Medicine
Nikolai-Fuchs-Straße 1, 30625 Hannover, Germany
{*christian.steinmeyer, †lena.wiese}@item.fraunhofer.de

***Abstract*—Along with digitization, automatic data-driven decision support systems become increasingly popular. Mortality prediction is a vital part of that decision process. With more data available, sophisticated machine learning models like (Artificial) Neural Networks (NNs) can be applied and promise favorable performance. We evaluate the reproducibility of a published mortality prediction approach using NNs along with the possibility to generalize it to a bigger and more generic dataset. We describe an extensive preprocessing pipeline, as well as the evaluation of different sampling techniques and NN architectures. Through training on a loss function that optimizes both, precision and recall, in combination with a good set of hyperparameters and a set of new features, we use a NN to predict in-hospital mortality with accuracy, sensitivity, and area under the receiver operating characteristic score of greater than 0.8.**

***Index Terms*—Medical information systems, Machine learning, Mortality prediction, Neural nets, Sampling**

***Funding*—This work was supported by the Fraunhofer Internal Programs under Grant No. Attract 042-601000.**

I. Introduction

Electronic Health Records (EHRs) are becoming ever more common in hospitals. As a result, so do automatic data-driven systems that support medical staff in their work. An important task due to its strong implications is mortality prediction. Here, we consider "*in-hospital-mortality*" especially in Intensive Care Units (ICUs), as its role in acute hospital care expands [1].

Not every staff member might know all indicators for a high risk condition and there might be some currently unknown complex cross-correlations between variables that are inconspicuous when looked at individually. In these cases, individual decision support is advantageous [2]. Automated prediction of mortality offers medical staff one more reference point for their decision making process.

The biggest challenges for working with EHRs and prediction tasks in a medical context are the heterogeneity, complexity and sparsity of the available data. Often, there are thousands of variables and measurements; yet, measurements might not available for all cases or samples [3]. And even if a variable is measured for multiple cases, it might be done differently for each. In [4], the authors analyze current databases for the *secondary use of EHRs*. They stress the challenges that arise from compartmentalization, corruption, and complexity, that naturally occur in the work context of intensive care. In a later publication, they show that much research lacks detailed descriptions of all of its proceeding which makes it difficult to reproduce [5]. In this article, we address this challenge by devising a workflow containing appropriate preprocessing steps.

One of the biggest and most comprehensive data sets containing ICU data is Multiparameter Intelligent Monitoring in Intensive Care - III (MIMIC-III). The MIMIC-III dataset is a large database containing a vast array of data regarding patients, measurements, survival data, and more [6, 7]. There are two main problems identified with this dataset with respect to mortality prediction: First, as the data collected in hospitals are very heterogeneous, they cannot be employed without preprocessing. Due to the sheer number of over 18000 different variables, their individual relevance to the prediction task is unclear. Second, the target variable of in-hospital-mortality is very imbalanced (with a ratio of approximately nine to one between survivors and fatal outcomes). Such an imbalance impacts the quality of many machine learning models like NNs [8]. However, in contrast to conventional machine learning techniques, a NN model might learn "*hidden*" correlations without the need of being taught explicitly, as they show particularly good performance in pattern recognition. Recent research motivates the use of many features, many data, and deep NNs [9, 10].

In this article, we demonstrate that the abundance of data in MIMIC-III allows the application of NNs. We show that the bigger dataset leads to better results when compared to the *PhysioNet* Challenge 2012 (PNC) dataset, although the latter was specifically created for the applied task of mortality prediction. We further evaluate the reproducibility of a related work, analyze different settings for simple NN architectures and balance strategies in order to reach satisfactory mortality prediction performance. By making use of (1) a small, targeted as well as (2) a bigger, generic dataset, we assess the generalizability of our findings.

The paper is outlined as follows. First, we discuss related work in Section II. Then we describe how we reproduced a NN approach to mortality prediction in ICUs by Chen and Yang in Section III. We use the resulting findings to evaluate simple sampling methods and NN architectures on the same dataset through a grid search in Section IV. We then apply the approach to a more generic dataset, MIMIC-III, and extend it through additional feature en-

gineering in Section V, and finally discuss our findings in Section VI.

II. Related Work

A. Mortality Prediction without Neural Networks

The de facto standards in the field of mortality prediction that are also used in some hospitals are predominantly two acuity scores: *Acute Physiology and Chronic Health Evaluation (APACHE)* I - IV [11, 12, 13, 14] and *Simplified Acute Physiology Score (SAPS)* I - 3 [15, 16, 17]. In 2012, the researchers published a challenge for mortality prediction using a specifically prepared subset of the dataset Multiparameter Intelligent Monitoring in Intensive Care - II (MIMIC-II) – which is the predecessor of MIMIC-III. They discuss how even this curated dataset incurs difficulties. The challenge results show that there are better methods for predicting mortality than the previously mentioned scores [18].

Some of the related research focuses on methodology. [19] performs personalized prediction through the analysis of only similar patients. They describe good results as long as the number of similar patients is big enough, but report higher performance cost compared to the standard approaches. [20] predicts sepsis as one of the most frequent causes of death at the ICU through machine learning on MIMIC-III. [21] suggests a solution on the PNC dataset through the use of a Bayesian ensemble learning model with 4000 primitive predictors. [22] performs unsupervised clustering of patients which leads to different mortality probabilities per cluster.

Other work focuses on data and features. [23] performs an automated feature selection "*[instead] of limiting the variables to those that have been validated to be predictive of mortality*" and yield good results. [24] performs feature construction from free text. They predict in-hospital-, 30-day-post-discharge, and 1-year-post-discharge-mortality. [25] analyzes just eight "*common clinical variables from the EHR*" and [26] uses only heart data which seems sufficient to achieve good results.

Still other work emphasizes the time of prediction. The norm is to predict mortality one day after admission, as this allows for some data to be gathered within the first 24 hours of an Intensive Care Unit Stay (ICU-stay). Sometimes, this observation period is also longer [3]. In contrast, [27] predicts mortality at time of admission. Similarly, [28] predicts mortality at six hours after admission.

As opposed to these approaches, we focus on predictions after 48 hours in the ICU. Features are selected through a combination of both, domain knowledge and statistical methods. Further, we employ an ensemble model for individual predictions.

Category	Options	[3]	[5]	[29]	Ours
Databases	PNC	✓	✓		✓
	MIMIC-III		✓	✓	✓
Feature Types	Medically motivated	✓	✓	✓	✓
	Other			✓	✓
Prediction Task	In-hospital mortality	✓	✓	✓	✓
	Other			✓	
Focus	Mortality Prediction	✓	✓	✓	✓
	Benchmarking			✓	
	Reproduction		✓	✓	✓
	Sampling				✓

TABLE I: Comparison of the most similar related works to ours.

B. Mortality Prediction with Neural Networks

In an early work, logistic regression and NNs were compared for mortality prediction in 2005 [30]. There, the authors conclude that in the majority of cases NNs outperformed logistic models in terms of discrimination and calibration, but not accuracy. The more accurate results for logistic models might be caused by a strong imbalance in the examined data.

Since then, NNs are used for a multitude of tasks, all with promising results and outperforming other approaches. [31] applies clustering on data with reduced dimensionality through *Auto-Encoders*. They further use *Long Short Term Memory (LSTM)* recurrent NNs, which consider temporal constellations of dynamic data, to predict 1-year-mortality. [32] uses *Auto-Encoders* to reduce dimensionality and enhance data with "*deep features*".

[33] analyzes unstructured text in patients' notes for mortality prediction on three levels (in-hospital, 30-day, 1-year). They report decent performance on MIMIC-III through convolutional NNs with a two-layered model for terms, and sentences. [34] applies self-normalizing NNs for this task. [35] predicts mortality through *deep* NNs with only few features that are widely available. [36] learns the task of a discharge decision. Other uses of deep learning for EHRs are surveyed in [37].

On the other side of the spectrum, [38] incorporates as many data into the prediction as possible. The authors perform a number of tasks. They predict mortality at multiple points in time and find very good performance of their model, an ensemble of three NNs. These are evaluated on two test sets coming from two distinct data sources. They reach compatibility by transforming the data into one single event stream of all information kinds per patient, mapping these events into a generic format, and analyzing them through their *deep* NNs. However, the approach required a lot of resources, both in terms of people that participated and processing power.

Finally, [29] provide an overview of additional approaches with NNs and create benchmarks for deep learn-

Fig. 1: Data processing pipeline with 6 processing steps to create the training data from group A of the raw data. The data's dimensionality is given after each step. The dimensionality term n refers to different lengths for different files, as the number of measurements taken for each patient varies.

ing (which are not applicable to our work due to differences in the pre-processing pipeline).

While NNs are already widely used for mortality prediction, many approaches do not report how they arrived at their model specification. Also, reproducibility and generalizability are limited. Thus, we aim to reproduce and extend an existing approach through the application towards a more generic dataset.

III. Reproduction of Chen and Yang's Work

In 2012, the *PhysioNet* team launched their annual challenge (https://www.physionet.org/content/challenge-2012) with a focus on mortality prediction for ICUs [39]. In a call for participation with the declared goal to "*develop methods for patient-specific prediction of in-hospital mortality*", a dataset, specifically engineered for this challenge, in combination with two scored events was published. This PNC dataset is generated from MIMIC-II. It is specifically designed for mortality prediction in ICUs and is equipped with a simpler structure than the original source.

As a baseline work [3] offers insights into the heterogeneity that underlies clinical data and provides a full processing pipeline from the challenge dataset to a NN model that performs well. In order to increase comparability to this prior approach, we reproduced their work closely with some useful alterations.

A. Data Processing

An overview of the processing pipeline is given in Figure 1. The processing steps are now explained in more detail:

Raw Data: The dataset contains 12000 ICU-stays, divided into three groups (A, provided with labels and used for training, B, provided for prediction task 1 and C, provided for prediction task 2) with data from the first two days after admission and up to 42 variables per stay. Out of the latter, five are general descriptors and 37 are time series; notably, weight is measured both as a general descriptor at admission, and as a time series throughout the stay. Each group contains 4000 files with n rows for different numbers of measurements and three columns each: *Timestamp (or Time)*, *Variable (or Parameter)*, and *Value*.

Note, that we use the same variable categories and definitions as introduced in [3], namely *General Descriptors*, *Low-sampling Variables*, *Med-sampling Variables*, and *High-sampling Variables*.

Merge Raw Data: We create one sparse matrix with data from all files of the group with individual measurements as rows and variables as columns. Each row only has three values: *Timestamp (or Time)*, *Case Identifier (or ID)*, and a value for one of the variables.

Remove Errors: We remove physiologically unrealistic entries for *Height* and *Weight* measurements (see Section III-B for details).

Apply Transformations: We calculate *Urine.Sum* from *Urine* and *CreatinineClearance* from *Creatinine*. We merge *TroponinI* and *TroponinT* to form a new variable *Troponin*, as well as invasive and non invasive blood pressure measurements, adding a binary variable that holds which method was used in the majority of measurements for each case.

Apply Imputations: We fill measurements that indicate missing values (because they are less than or equal to zero or outside the meaningful range of variable values) by the variable category dependent imputation method suggested in [3], either gender specific mean values or random draws from a gender specific normal distribution for that variable. In the same way, completely absent data are imputed, that is, if no measurement is available at all for a given case for a given variable.

Extract Features: Depending on the variable category, we apply the according feature extraction methods from [3]: leave *General Descriptors* unchanged and turn *Low-Sampling Variables* into a categorical feature stating either that the variable was not measured for a given case, that all measured values are normal or at least one value was abnormal. For *Med-Sampling Variables*, use the mean of all variable measurements for a given case and for *High-Sampling variables*, also provide their minimum, maximum, median, first measurement, last measurement, linear trend over the first 24 hours, linear trend over the second 24 hours, and linear trend over 48 hours (yielding 9 summary statistics). In total, we extract the same 112 features from the PNC dataset as in [3].

Select Features: Using the feature selection method *Minimum Redundancy Maximum Relevance*, the most rel-

evant features are identified and the others disregarded.

In [3], they define a not further justified relevance threshold of -0.005 for the *mRMR* score, yielding 47 features, only the top five of which are known: *GCS_last*, *BUN_mean*, *PaCO2*, *Platelets_mean*, and *SABP_mean*[1]. Note, that we found similar, but different results: we found the top five to be *GCS_last*, *CreatinineClearance*, *Platelets*, *SABP_mean*, and *PaCO2* (for the features we use, see Section III-B).

B. Deviations and Assumptions

During the preprocessing stage, we deviate from [3] in the following ways or make the following assumptions, if no information was given:

- weight measurements with less than 10 kilograms are removed, as only adult patients are observed
- height measurements with less than 100 or more than 300 centimeters are removed
- creatinine transformation is performed using the most recent weight or the gender based mean weight, if none is available
- imputation method *One* is interpreted as choosing the gender-specific mean values
- imputation method *Four* is interpreted as randomly drawing from the normal distribution for each variable defined by its mean and standard deviation
- all underlying mean and standard deviation values are calculated from the designated training dataset and the same values are used for the test datasets
- trends with too few observations (less than 10 observations in 48 hours or 5 observations in either 24-hour period) receive a trend value of zero

For the feature selection, in addition to the *Minimum Redundancy, Maximum Relevance* method, we apply *Chi-Square (Chi²)* due to its wide-spread use and availability in popular libraries, which allows for easier integration in workflows and yields less ambiguous scoring results. This feature selection method is based on the *chi²* statistical test that evaluates the dependence between the target variable and a feature. It requires all values to be positive, which is why we normalize the data to be between zero and one prior to the feature selection. A side benefit of this is that NNs converge faster as they do not need to learn different scales for different features [40]. Following the *Elbow-Method*, we determine $n = 20$ for getting the most useful information by the least number of features (cf. Figure 2). We found similar results for both feature selection methods and thus continue with the fewer features selected through *Chi²*.

[1]The feature names deviate slightly from [3] for a better fit in this work: e. g., what we call *GCS_last*, they call *GCS.last_input*.

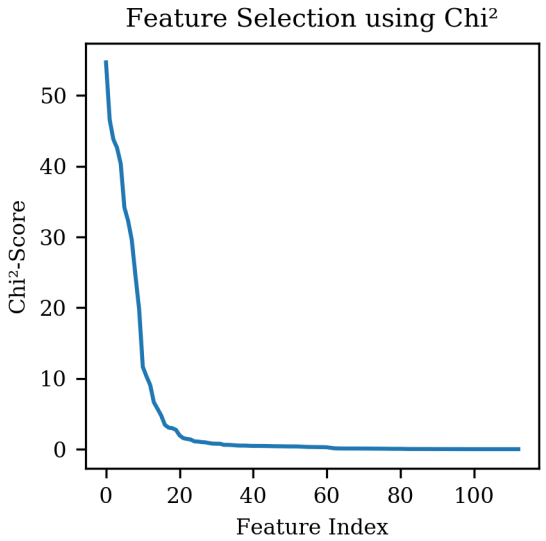

Fig. 2: The sorted *Chi²* scores for all features.

Even though reproducing the pipeline from [3] as closely as possible, we could not reproduce the suggested model's results – most likely due to multiple ambiguities in their description or random processes that are part of the data processing pipeline. Especially, the *log-sigmoid* activation function in the output layer of the NN yielded sole prediction of the majority class despite class balancing. Thus, we make the following changes:

- *softmax* is used as activation function in the output layer [41]
- *relu* is used as activation function in the hidden layers [42]
- *adam* is used as optimizer [43]
- weights are initialized using random values from a normal distribution

Furthermore, as no information was provided in [3] regarding ensemble size and loss function, we define them as seven and *F1*, the harmonic average of precision and recall, respectively [44]. These changes yield a model with better predictive power than the original one; we investigate its optimal configuration next.

IV. Evaluation of Simple Sampling Methods and Neural Network Architectures

For the implementation we used *Python*, *NumPy*, *pandas*, *scikit-learn*, *imbalanced-learn*, *Jupyter Notebook*, *Keras*, and *TensorFlow*. In order to enable reproduction of our full results, all source code used for this work is published in a publicly available repository: https://gitlab.com/christian-steinmeyer/nnmp.

We systematically evaluate different settings by means of a grid search in order to determine the best combination of NN architectures and sampling methods for the given data. To that end, we perform 6720 experiments in 336 different settings and compare their results in the following way.

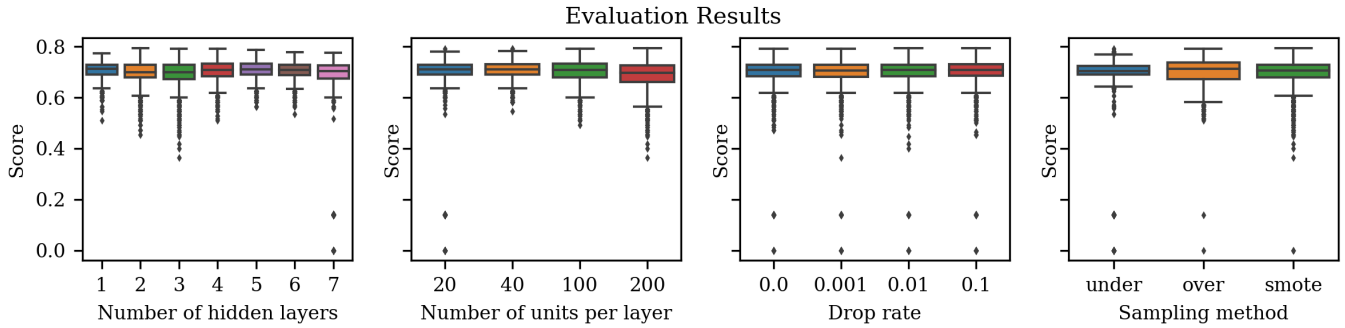

Fig. 3: Evaluation results for the PNC dataset displaying the scores by number of hidden layers ($n = 960$), number of units per layer ($n = 1680$), drop rate ($n = 1680$), and sampling method ($n = 2240$) from left to right.

A. Methods

In terms of classification we define the fatal outcome of a patient as a positive sample, the non-fatal outcome as a negative sample. Due to the sensitive nature of mortality prediction as well as the target class imbalance in the given data, sensitivity (or recall) is considered in addition to the common binary classification metrics accuracy and Area Under the Receiver Operating Characteristic Curve (AUROC). Additionally, we use a score, defined as the minimum of sensitivity, accuracy, and AUROC, as evaluation metric for this work. This gives more weight to the False Negatives (FN), than just using accuracy or AUROC. FN denote the incorrect prediction of survival for a fatal outcome, which is more important in this context in order to not overlook patients who are indeed in severe risk of death. Note, that we use *Keras*' classification threshold of 0.5 for predictions.

In order to receive reliable results despite the random processes involved in the model, as well as the target class imbalance, we use a suitable variant of *cross validation*: We split the available data randomly into k equally sized "*folds*" (or parts), each of which is used as test data once while the combination of the other $(k-1)$ folds are used as training data. Thus, in k iterations, k model instances are trained and evaluated on the available data. This process is called cross validation [45]. A cross validation is called stratified, if the target class imbalance is retained in each fold. We use a repeated version with different splits in each iteration, called repeated stratified k-fold cross validation. This method also helps to avoid over-fitting.

B. Experiments

We consider "*group A*" from the PNC dataset, that is 4,000 cases, as training data, preprocessed in the manner described above. We evaluate different architectures of simple NNs and balance strategies (i. e., sampling methods) through a grid search approach. For that, we consider all possible combinations or settings (d, w, dr, ba) from the following parameters:

- network depth d: number of hidden layers (d in $[1, 2, 3, 4, 5, 6, 7]$)
- network width w: number of units per layer (w in $[20, 40, 100, 200]$)
- drop rate dr: rate of data dropped per layer (dr in $[0, 0.001, 0.01, 0.1]$)
- balance strategy ba: one of random under-sampling, random over-sampling, or Synthetic Minority Over-Sampling Technique (SMOTE)

To achieve reliable results, we apply repeated stratified cross validation as explained in Section IV-A with two repetitions and ten folds each, yielding 20 separate runs for each of the 336 settings.

C. Results

The results of the grid search are depicted in Figure 3. There, for each of the observed parameters, the resulting scores for each value from above experiments are given in a boxplot. Each score corresponds to one iteration of the cross-validation on "*group A*". From their analysis, we formulate the following concluding theories: Deeper networks also require a certain width to cope with their complexity , as the worst results are achieved with seven hidden layers and 20 units per layer, independent of drop rate and sampling method. With an increase of units per layer, the results are less consistent (the result range becomes bigger). Drop rate has little effect on the score of the model in this context. There is not only important information in the minority class data, but also in the majority class data as well.

Hyperparameter search was done using group A for training and group B for testing. The best setting or combination of parameters is a model using five hidden layers, 20 units per layer, a drop rate of 0.001 and over-sampling. When trained on group A of the PNC dataset and evaluated on group C this model yields an accuracy of 0.726, AUROC score of 0.749, sensitivity of 0.783, specificity of 0.716, and therefore an overall score of 0.726. Moreover, if the model is trained on both, group A and B and tested on group C, it yields these metrics: accuracy 0.748, AUROC score 0.756, sensitivity 0.768, and overall score 0.748. In comparison, Yun Chen and Hui Yang report for their model an accuracy of 0.91, AUROC score of 0.88, sensitivity of 0.83, and therefore an overall score of 0.83.

Hence we unfortunately were unable to reproduce their reported results despite recreating their settings as close as possible to the best of our knowledge.

V. Application Towards MIMIC-III

While the PNC dataset underwent special preparation before publication and is tuned for mortality prediction, most real world scenarios do not offer this grade of specialization. The MIMIC-III dataset for example contains a vast array of data that goes beyond patient data, relevant measurements, and survival data [6, 7]. The dataset consists of 26 tables with up to 330 million entries (we use MIMIC-III version 1.4 from September 2016). As it is designed to support a diverse range of analytics, it is not tailored towards mortality prediction.

Not only is the MIMIC-III dataset much bigger in comparison to the PNC dataset, but it also includes much more diverse data. In order to remain maintainable and to comply with a multitude of use cases, it is structured differently – in particular, it is more fragmented. This requires access to multiple tables including complex queries to a database for a full set of patient-level information. For the sake of mortality prediction, we only consider nine tables containing information about admissions, (measurement) events, item definitions, ICU-stays, patients, and transfers.

In the remainder of this section, the different steps of our data analysis process are described.

A. Exploration

We made the following observations when exploring the MIMIC-III dataset. The target class is also imbalanced here, even slightly stronger than in the PNC data with 9 fatalities out of 100 ICU-stays. MIMIC-III contains wrong data (i. e., data that are formally correct but not plausible, like 4000 breaths per minute) and missing data (i. e., some measurements contain values less than or equal to zero for variables where this is outside the reasonable range) as well as absent data (not all patients were subject to all measurements). Further, MIMIC-III is curated from multiple data sources which leads to some ambiguities like duplication or inconsistent units. The data are overall as heterogeneous as in PNC, if not more so. Lastly, we use the criteria introduced in [3] for variable categorization and apply them to MIMIC-III at a later time (see Section V-C).

B. Preprocessing

Due to the size difference of multiple orders of magnitude of this dataset, we remove redundant and irrelevant data to reduce memory usage and run time; we define irrelevant as not being part of investigation like ICU-stays that are shorter than two days. We consider all care units except NWARD and NICU. We exclude patients and additional data that do not comply to the above requirements, yielding overall 69077 transfers, 35446 patients (57% male), 46130 admissions (28984 patients with single admission), 563982 diagnoses, 49059 ICU-stays, 23205670 lab events, and 35076620 chart events.

Then, we select the variables given in the PNC dataset. Because MIMIC-III has multiple sources, measurements of the same type exist, but might have different identifiers or units. We join them manually, converting all data to the same unit (e. g., if multiple identifiers exist for the measurement type *weight*, some might be in *kilograms* and others in *pounds*). Variable transformations are performed analogously to before, with the exception of blood pressures, as MIMIC-III differs less strictly between invasive and non invasive methods.

As many more variables are available in MIMIC-III, we also select the 99 most frequent lab measurements, defined as appearing in the most hospital admissions. These measurements should be easy and inexpensive as they are taken frequently and yet contain useful information for the prediction of mortality as we show later in Section V-F. We neglect all lab and chart events with fewer than 100 numeric measurement values across all relevant hospital admissions.

Again, erroneous data are removed, i. e., data outside of realistic ranges, defined through statistically motivated bounds given in Table II. They are defined by (inclusive) lower and upper thresholds. Excluding zero is only relevant if the lower bound is a positive number. Zero values indicate missing data and are later on replaced instead of simply being removed.

In order to split the data into training and test sets, two thirds are selected in a shuffled and stratified manner (maintaining the imbalance ratio) for training data. One third is held back for testing only.

Variable	Lower Bound	Upper Bound	Excl. Zero
Amylase		7000	
Base Excess	-100	100	
Creatinine [mg/dL]		200	
Diastolic ABP [mmHg]		10000	
Epithelial Cells		400	
FiO2		2000	
Glucose [mg/dL]	0		
Heart Rate [bpm]	0	400	
Height [cm]	50	300	Yes
Hematocrit (HCT) [%]		100	
Hemoglobin		40	
Lipase		200000	
Oxygen	0	500	
Oxygen Saturation in Hemoglobin (SaO2) [%]	0	100	
Potassium (K)		100	
PT		150	
RDW	10		Yes
Respiratory Rate [bpm]	0	500	
Sodium (Na)		200	
Sodium Urine		300	
Systolic ABP [mmHg]		10000	
Specific Gravity	1		Yes
Temperature (Temp) [°C]	0	50	
Weight [kg]	10	1000	Yes

TABLE II: Unrealistic Variable Thresholds with units as given in MIMIC-III

C. Data Imputation

We handle all numeric time series data like high-sampling variables, also regarding data imputation. Thus, we impute missing values through gender-specific normal distributions, the necessary mean and standard deviation values are created from the available data in the training set. In a first step, all measurements with a value of zero (where zero is not part of the normal range) are replaced. In a second one, new synthetic measurements are added, if none exist for a certain variable for a given case. As the missing rates of variables are now known, we can determine the variable categories at this point.

D. Feature Extraction and Selection

Since we treat all numeric time series data like high-sampling variables, we create static features through the same nine summary statistics as before (see Section III), with an additional trend value for measurements taken before a given admission (e. g., during prior visits). When applying Chi^2 in order to reduce the number of features, a trade-off between informativeness and efficiency is identified for 100 features. Therefore, only those 100 features are selected as our final dataset.

E. Comparing Evaluation Between the Two Datasets

Before comparing the evaluation results, we inspect differences in feature selection and therefore our versions of the datasets that are used within the evaluation.

When comparing the selected features from the dataset to the ones selected from the MIMIC-III dataset, the following observations can be made: Out of the Top 20 important features in PNC, 13 are also identified as one of the Top 100 important features in MIMIC-III (either once if the corresponding variable was categorized as high-sampling, or even multiple times if it was a low- or med-sampling variable like *Creatinine Clearance*), five are not selected (namely, *Troponin*, *ALP*, *SaO2*, *GCS_trend_day_one*, and *GCS_trend_two_days*), and two features are not available in the same manner and therefore not comparable. The top 100 features in MIMIC-III are created from 33 different variables. Out of those, 21 were also part of the PNC. However, the following other twelve features are identified as important and highly available and therefore potentially useful for mortality prediction: *Anion Gap*, *Bands*, *Base Excess*, *INR(PT)*, *Lactate Dehydrogenase (LD)*, *Lipase*, *Lymphocytes*, *MCHC*, *Phosphate*, *PTT*, *PT*, and *RDW24*.

To the end of comparing the two datasets from PNC and MIMIC-III, we apply the same methods and experiments as described in Section IV with the only difference of a constant drop rate of zero as its variation showed no effect for the PNC dataset. Thus, we perform a another exhaustive search with a total of 1680 experiments in 84 different settings, automatically executing all possible combinations of the above parameters.

The results of the grid search are depicted in Figure 4. There, for each of the observed parameters, the resulting scores for each value are given in a boxplot. From their analysis, we formulate the following concluding theories: Not only require deeper networks a certain width to cope with their complexity, but there seems to be an upper limit to that width, which depends on the input dimensions or data. Higher numbers of hidden layers (network depth) lead to increasingly unstable results (wider range) and worse performance (in regard to worst and average cases, while the best case is not affected). A similar tendency is observed for the number of units per layer (network width). With more features being selected in MIMIC-III, compared to PNC, sampling methods that generate new data samples based on statistical relationships have more chances for statistical errors which leads to poorer performance. The problem complexity when using 20 features is better manageable with the set of models evaluated in this work and the amount of training data, while using 100 features would require more training data for more consistent results.

The best setting when restricting to the Top 100 features is a model using one hidden layer, 200 units per layer, and under-sampling. When trained on the whole set of training data (two thirds of the data) from MIMIC-III and evaluated on the test set, this model yields an accuracy

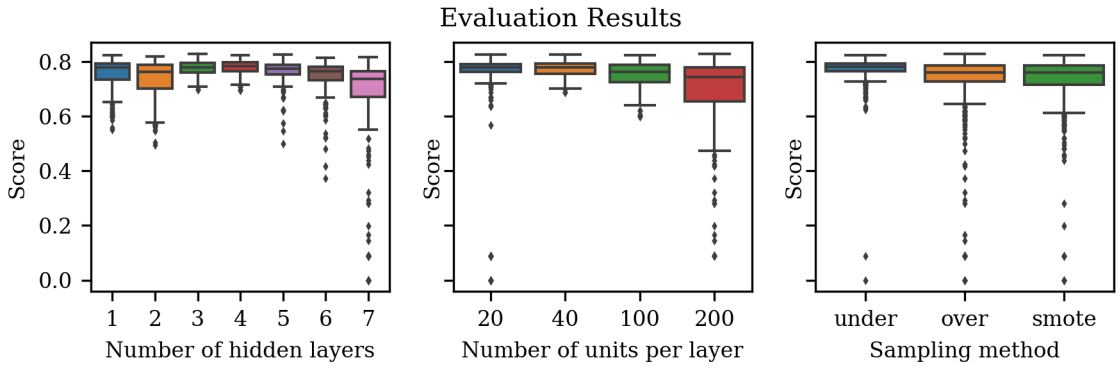

Fig. 4: Evaluation results for the MIMIC-III dataset displaying the scores by number of hidden layers, number of units per layer, and sampling method (from left to right).

of 0.778, AUROC score of 0.805, sensitivity of 0.836, and therefore an overall score of 0.778. Yet, when using all 821 features, the model yields an accuracy of 0.811, AUROC score of 0.810, sensitivity of 0.809, and therefore an overall score of 0.809 at the cost of a significantly increased training time.

F. Additional Feature Engineering

Since the MIMIC-III dataset offers more information than the PNC dataset, some of those additional data might be helpful for mortality prediction. In some cases it contains data from previous stays and in most cases diagnoses. We engineer several additional features and evaluate them as described below. All of them are created on a per ICU-stay basis.

Previous stays and their data as well as event data from the current ICU-stay are used to engineer the following five features. The intuition behind them being that patients who were severely ill before, might be weakened by their previous illnesses or overall be more susceptible to diseases or to other strong consequences from health issues. As another hypothesis, there exists a correlation with the number of measurements that are taken for a patient and the severity of the patient's illness.

- *Number of Previous Stays*, defined as the number of ICU-stays with the same *SUBJECT_ID* that occurred before the one in question.
- *Number of Previous Events*, defined as the number of (lab or chart) events that were measured before the beginning of this case.
- *Number of Abnormal Previous Events*, defined as the number of (lab and chart) events that were measured before the beginning of this case, but yielded an abnormal result.
- *Number of Observation Period Events*, defined as the number of (lab or chart) events that were measured within two days after the beginning of this case (as the observation period was defined earlier to be 48 hours).
- *Number of Abnormal Observation Period Events*, defined as the number of (lab or chart) events that were measured within two days after the beginning of this case, but yielded an abnormal result.

Diagnosis data from previous admissions are used to engineer the following features. The intuition behind them being that diagnoses carry a lot of information that severely influence the treatment of patients and possibly also their outcome. ICD9 codes are organized in 19 groups or clusters of diagnoses.

- *Binary ICD9 Group Features*, defined for each diagnosis group as 1, if at least one diagnosis from that group was made for the given patient and the hospital admission associated with this case, and as 0 otherwise. Note, that there is no time information available for the diagnoses in MIMIC-III, which makes it impossible to determine which diagnosis was made before or after the observation period. In using these data nonetheless, we assume, that usually diagnoses are made early within hospital stays. This yields 19 features.
- *Diagnosis Based Mortality Rates*, defined as the minimum, mean, or maximum of the ICD9 group mortality rates for all groups where the correlating binary ICD9 group features for this case from above equal 1. The ICD9 group mortality rates are determined beforehand through the training set. For each group, the total number of cases with a corresponding binary ICD9 group feature equal to 1 (n) and out of those the number of cases with an in-hospital mortal outcome are determined (m). The mortality rate (r) of that group is simply defined as $r = \frac{m}{n}$. This yields three features.

Overall, we create 27 newly engineered features, adding to a new total of 848 features, when combined with the existing ones from Section V. When applying Chi^2 feature selection to select the Top 100 most relevant features like before, out of the newly engineered features, twelve are included.

To evaluate the effects of these features, the best performing setting and model from Section V-E (one hidden layer, 200 units per layer, and under-sampling) is trained as before on the whole set of training data and evaluated

on the test set, including the new features. Now, this model yields an accuracy of 0.801, AUROC score of 0.804, sensitivity of 0.808, and therefore an **overall score of 0.801**. For reference, when using all 848 features including the new ones, the model yields an accuracy of 0.819, AUROC score of 0.822, sensitivity of 0.826, and therefore an overall score of 0.819.

VI. Discussion

A. Findings

Not only can the approach developed in Section III for the PNC dataset be applied to the MIMIC-III dataset, but the prediction performance actually improves when doing so. The MIMIC-III dataset was found to contain the same difficulties with respect to data heterogeneity as the PNC dataset. Additionally, it combines multiple data sources and thus, it contains ambiguities as well as redundancies. As it offers a lot more data, the former requires data reduction as additional preprocessing step. At the same time, it offers the potential to consider more variables and therefore features of interest. In addition to the variables contained in the PNC, we also evaluate the most available variables in ICUs. We could not apply all data transformations from the PNC dataset to the MIMIC-III dataset because the underlying information and structure was different. Due to those differences, we decided to not use the data categorization from [3], but to treat all variables as high-sampling variables. This led to a higher number of features, the analysis of which through the Chi^2 method showed that 100 seem relevant to in-hospital mortality. We found that measurements available for most ICU-stays are helpful for mortality prediction, as are diagnosis data and meta features.

In all evaluated cases, an imbalanced dataset without sampling resulted in the sole prediction of the majority class. The results regarding different simple feed forward NN architectures and random under-, over- and SMOTE-sampling methods with respect to mortality prediction differ from dataset to dataset. The evaluation of simple feed forward architectures defined by the number of hidden layers (one through seven were evaluated), the number of units per layer (20, 40, 100, 200), and by the drop rate used before each hidden layer (0, 0.001, 0.01, 0.1) allow the following conclusions: We evaluated drop rate only for the PNC dataset but observed no significant influence. Deeper networks require a certain width to cope with their complexity. With an increase of units per layer, the results are less consistent. In the PNC setting, slim networks with less than 100 units per layer tend to perform best and there is not only important information in the minority class data, but also in the majority class data, as over-sampling techniques have better peak performance. Under-sampling yields more consistent results than over-sampling or SMOTE. Balancing is effective for mortality prediction with NNs.

Differences in performance between very different models are small in all cases. Thus, our best performing settings do not depict widely generalizable configurations, but rather, dataset specific maxima. Overall, NNs seem to be able to learn mortality prediction from clinical data. Despite this limitation, we can make more general deductions that are applicable for a wider range of datasets. With increased input dimensions, a higher number of units per layer decreases stability, as well as worst case and average case results of the network's prediction capabilities. The same observation can be made for higher numbers of hidden layers. While, in terms of accuracy, too much complexity (many hidden layers with many numbers per layer and over-sampling methods) leads to increasingly bad results, less complexity allows for better results. There exists a tendency for the tested NNs to have better accuracy than sensitivity. The best scoring models have more True Positives (TP), but they predict fatal outcomes overall more often, which leads to an increase in False Positives (FP) and therefore also in accuracy, which is preferable over FN.

Selecting variables based on domain knowledge, as was done for the PNC dataset, enables the prediction of mortality with satisfactory results. However, not all of these features seem to be necessary for a successful prediction (see Section III). Additionally, there exist variables, that are widely available (partially much more available than the former) and also contain helpful information for mortality prediction. We identified *Anion Gap*, *Bands*, *Base Excess*, *INR(PT)*, *Lactate Dehydrogenase (LD)*, *Lipase*, *Lymphocytes*, *MCHC*, *Phosphate*, *PTT*, *PT*, and *RDW*. Creating static features from time series variables through data aggregation and the use of summary statistics, the statistical feature selection method Chi^2 suggests, that the most important statistics are the last and maximum values measured during the observation period. In contrast to the intuition that time series analysis through dynamic features and therefore without information loss would perform favorably, it was observed, that linear trend information from the available data is less relevant to the successful prediction of in-hospital mortality in our approach. We found that feature engineering can further improve the prediction results: We engineered 27 features from previous stay data, an analysis of the amount of measurements taken within the observational period and diagnoses. Out of those, 12 were identified to be more relevant than the main portion of medically motivated features, given in decreasing order by Chi^2 score: *Diseases of the Respiratory System*, *Infectious and Parasitic Diseases*, *Symptoms, Signs, and Ill-Defined Conditions*, *Diseases of the Genitourinary System*, *Number of Abnormal Observation Period Events*, *Mental Disorders*, *Injury and Poisoning*, *Neoplasms*, *Supplementary Classification of External Causes of Injury and Poisoning*, *Diseases of the Musculoskeletal System and Connective Tissue*, *Diseases of the Digestive System*, and *Congenital Anomalies*. Note, that except of *Number of Abnormal Observation Period Events*, these refer to ICD9 groups and whether the case under consideration contains a diagnosis belonging to the corresponding ICD9 group or not.

B. Future Work

In future work, further improvements can be made to the approach. They include, but are not limited to: Evaluation of different classification thresholds (cf. Section IV-A).

C. Conclusion

We show, that it is possible to reproduce the research approach published by Yun Chen and Hui Yang for mortality prediction on the PNC dataset, if the model is adjusted and some assumptions about data processing are made, to a certain extent. But the reported performance of their model could not be fully reached. The adjusted approach could, however, be generalized to a broader data set, namely MIMIC-III and achieved decent performance. That is, to predict in-hospital-mortality to an accurate degree using NNs and according preprocessing of the clinical data available up until the first 48 hours after a transfer into the ICU as well as a target class balance strategy. Challenges in data heterogeneity can be overcome by using highly available variables, additionally to medically motivated variables, and creating static features through summary statistics. A simple feed forward network with one hidden layer and 200 units in that layer suffices to achieve accuracy, sensitivity and AUROC scores of greater than 0.8 for MIMIC-III. To reach an efficient and effective training process, we identified a balanced and reasonably preprocessed dataset as the most important factors. Through training on a loss function that optimizes the *F-score*, it is possible to maintain high sensitivity, accuracy and AUROC scores at the same time. Both, the effect of sampling methods and the architecture of the NN on the results depend on the dataset. Drop rate had no beneficial effect when combined with an ensemble voting model and cross validation. Further, with an increase of units per layer and of the number of hidden layers, network results become less consistent, especially for bigger input dimensions. Finally, we improved the performance of the approach through the engineering of new features.

References

[1] M. Ghassemi, L. A. Celi, and D. J. Stone, "State of the art review: The data revolution in critical care," *Critical Care*, vol. 19, no. 1, p. 118, 3 2015.

[2] L. Celi, R. Tang, M. Villarroel, G. Davidzon, W. Lester, and H. Chueh, "A Clinical Database-Driven Approach to Decision Support: Predicting Mortality Among Patients with Acute Kidney Injury," *Journal of Healthcare Engineering*, vol. 2, no. 1, pp. 97–110, 3 2011.

[3] Yun Chen and Hui Yang, "Heterogeneous postsurgical data analytics for predictive modeling of mortality risks in intensive care units," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2014. IEEE, 8 2014, pp. 4310–4314.

[4] A. E. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. Clifton, and G. D. Clifford, "Machine Learning and Decision Support in Critical Care," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444–466, 2 2016.

[5] A. E. W. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: a mortality prediction case study," *Machine Learning for Healthcare Conference*, vol. 68, pp. 361–376, 2017.

[6] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 5 2016.

[7] A. E. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "The MIMIC Code Repository: Enabling reproducibility in critical care research," *Journal of the American Medical Informatics Association*, vol. 25, no. 1, pp. 32–39, 1 2018.

[8] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2-3, pp. 427–436, 2008.

[9] M. Bhandari and M. Reddiboina, "Building artificial intelligence-based personalized predictive models," *BJU international*, 3 2019.

[10] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P. M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. Decaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, and C. S. Greene, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of the Royal Society Interface*, vol. 15, no. 141, p. 20170387, 4 2018.

[11] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, "APACHE-acute physiology and chronic health evaluation: a physiologically based classification system," *Critical care medicine*, vol. 9, no. 8, pp. 591–7, 8 1981.

[12] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: a severity of disease classification system," *Critical care medicine*, vol. 13, no. 10, pp. 818–29, 10 1985.

[13] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano, "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–36, 12 1991.

[14] 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," *Critical*

Care Medicine, vol. 34, no. 5, pp. 1297–1310, 5 2006.

[15] J. R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers, "A simplified acute physiology score for ICU patients," *Critical care medicine*, vol. 12, no. 11, pp. 975–7, 11 1984.

[16] J. R. Le 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. 2957–63, 1993.

[17] P. G. Metnitz, R. P. Moreno, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, and J. R. Le Gall, "SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description," *Intensive Care Medicine*, vol. 31, no. 10, pp. 1336–1344, 10 2005.

[18] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012," *Computing in cardiology*, vol. 39, pp. 245–248, 2012.

[19] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PLoS ONE*, vol. 10, no. 5, p. e0127428, 5 2015.

[20] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, and R. Das, "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach," *JMIR Medical Informatics*, vol. 4, no. 3, p. e28, 9 2016.

[21] A. E. W. Johnson, N. Dunkley, L. Mayaud, A. Tsanas, A. a. Kramer, and D. Clifford, "Patient Specific Predictions in the Intensive Care Unit Using a Bayesian Ensemble," *Computing in Cardiology*, no. Mimic, pp. 249–252, 2012.

[22] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*, 2012, p. 389.

[23] L. A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, and R. Mark, "A database-driven decision support system: Customized mortality prediction," *Journal of Personalized Medicine*, vol. 2, no. 4, pp. 138–148, 9 2012.

[24] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding Physiological State: Mortality Modelling in Intensive Care Units," *International Conference on Knowledge Discovery & Data Mining*, vol. 2014, pp. 75–84, 8 2014.

[25] J. Calvert, Q. Mao, J. L. Hoffman, M. Jay, T. Desautels, H. Mohamadlou, U. Chettipally, and R. Das, "Using electronic health record collected clinical variables to predict medical intensive care unit mortality," *Annals of Medicine and Surgery*, vol. 11, pp. 52–57, 11 2016.

[26] R. Sadeghi, T. Banerjee, and W. Romine, "Early hospital mortality prediction using vital signals," *Smart Health*, vol. 9-10, pp. 265–274, 12 2018.

[27] N. Veith and R. Steele, "Machine Learning-based Prediction of ICU Patient Mortality at Time of Admission," in *Proceedings of the 2nd International Conference on Information System and Data Mining - ICISDM '18*. New York, New York, USA: ACM Press, 2018, pp. 34–38.

[28] A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs, "Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach," *International Journal of Medical Informatics*, vol. 108, pp. 185–195, 12 2017.

[29] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *Journal of biomedical informatics*, vol. 83, pp. 112–134, 2018.

[30] B. Eftekhar, K. Mohammad, H. E. Ardebili, M. Ghodsi, and E. Ketabchi, "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data," *BMC Medical Informatics and Decision Making*, vol. 5, no. 1, p. 3, 12 2005.

[31] B. K. Beaulieu-Jones, P. Orzechowski, and J. H. Moore, "Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database," in *Biocomputing 2018*. World Scientific, 2 2017, pp. 123–132.

[32] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Scientific Reports*, vol. 6, no. 1, p. 26094, 9 2016.

[33] P. Grnarova, F. Schmidt, S. L. Hyland, and C. Eickhoff, "Neural Document Embeddings for Intensive Care Patient Mortality Prediction," *CoRR*, vol. abs/1612.0, 12 2016.

[34] M. A. Zahid and J. Lee, "Mortality prediction with self normalizing neural networks in intensive care unit patients," in *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018*, vol. 2018-Janua. IEEE, 3 2018, pp. 226–229.

[35] H. R. Darabi, D. Tsinis, K. Zecchini, W. F. Whitcomb, and A. Liss, "Forecasting mortality risk for patients admitted to intensive care units using machine learning," *Procedia Computer Science*, vol. 140, pp. 306–313, 2018.

[36] C. J. McWilliams, D. J. Lawson, R. Santos-Rodriguez, I. D. Gilchrist, A. Champneys, T. H. Gould, M. J. Thomas, and C. P. Bourdeaux, "Towards a decision support tool for intensive care discharge: Machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol,

UK," *BMJ Open*, vol. 9, no. 3, p. e025925, 3 2019.

[37] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[38] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, P. J. Liu, X. Liu, M. Sun, P. Sundberg, H. Yee, K. Zhang, G. E. Duggan, G. Flores, M. Hardt, J. Irvine, Q. Le, K. Litsch, J. Marcus, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. Howell, C. Cui, G. Corrado, and J. Dean, "Scalable and accurate deep learning for electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 18, 12 2018.

[39] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. 215–20, 6 2000.

[40] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," *CoRR*, vol. abs/1706.0, 6 2017.

[41] C. Bishop, *Neural networks for pattern recognition*. Clarendon Press, 1996.

[42] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807 – 814.

[43] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 12 2014.

[44] C. V. Rijsbergen, "Information retrieval," 1979.

[45] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1 1974.