Knowledge Distillation from Machine Learning Models for Prediction of Hemodialysis Outcomes

Harry Freitas da Cruz, Siegfried Horschig and
Matthieu-P. Schapranow

Digital Health Center
Hasso Plattner Institute
Rudolf-Breitscheid-Str. 187, 14482 Potsdam, Germany
Email: {harry.freitasdacruz|schapranow}@hpi.de
siegfried.horschig@student.hpi.de

Christian Nusshag

Department of Nephrology
Heidelberg University Hospital
Im Neuenheimer Feld 162, 69120 Heidelberg, Germany
Email: christian.nusshag@med.uni-heidelberg.de

***Abstract*—In order to compensate severe impairments of renal function, artificial, extracorporeal devices, so called dialyzers, have been developed to enable renal replacement therapy. The parameters utilized in this form of therapy and the specific patient characteristics substantially affect individual patient outcomes and overall disease progression. In this paper, we present a clinical prediction model for outcomes of critically ill patients that underwent a specific form of renal replacement, hemodialysis. For this purpose, we employed two categories of machine learning models: interpretable (Bayesian rule lists and logistic regression) and non-interpretable (multilayer perceptron and random forest). To provide more transparency to the latter category, we applied mimic learning and feature importance metrics. Results show that non-interpretable models outperform the rule-based classifier (c-statistic $\geq$ 0.9). Despite this result, the use of interpretability methods enables more thorough model scrutiny by a medical experts, revealing possible model biases, which might have been otherwise disregarded.**

Keywords*–clinical prediction model; renal replacement therapy; machine learning; supervised learning; knowledge distillation.

I. Introduction

Previously, we developed a prediction model for patient outcomes following Renal Replacement Therapy (RRT) [1]. In this paper, we expand our previous work, including different algorithms, metrics, and more in-depth discussion so as to provide a more comprehensive picture of the contributions, challenges and limitations faced.

The renal system in the human body has the purpose to excrete predominantly water-soluble metabolites and toxins in order to maintain a sufficient blood homeostasis [2]. If this system is impaired severely, e.g., in the context of an Acute Kidney Injury (AKI), artificial, extracorporeal organ replacement therapy becomes necessary [3]. Therefore, different RRT modalities are available. One example is the hemodialysis, where the solute exchange takes place via diffusion across a semipermeable membrane between the blood and the dialysate or dialysis fluid [4].

Hemodialysis outcomes are highly dependent on the patient's clinical characteristics as well as on the type of the RRT procedure applied [5]. Furthermore, RRT modalities based on a filtration circuit, such as hemofiltration or hemodiafiltration are particularly costly, requiring specialized equipment and nursing staff [6]. In addition, various parameters have to be adjusted for each patient, e.g., duration of the process, the filtration rate and flow rates of the blood and dialysate. Clinical prediction models can aid in decision making by providing nephrologists with more accurate prognostic information under uncertainty of outcomes [7].

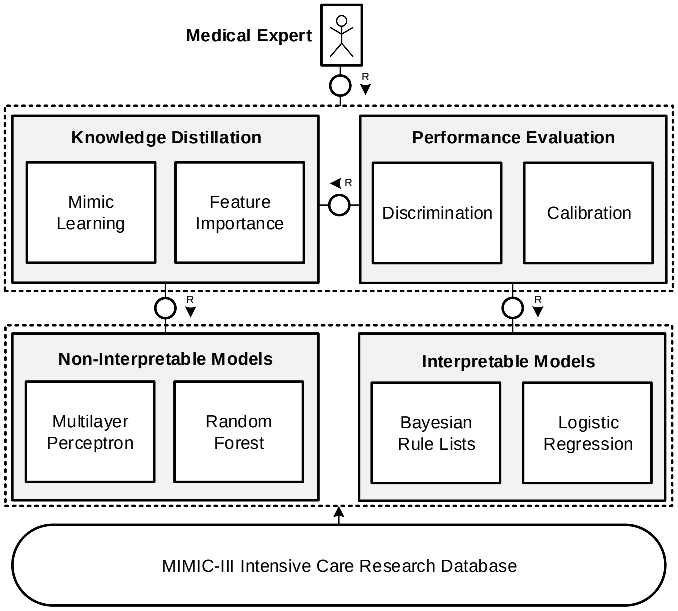


Figure 1. Our research setup modeled as a Fundamental Modeling Concepts (FMC) block diagram. Knowledge distillation approaches allow a medical expert to scrutinize the non-interpretable, black-box models.

In addition to usual criteria like accuracy or recall, when employing Machine Learning (ML) in the medical context, one especially important factor is the interpretability of the model, since doctors must take full responsibility for the respective decision, therefore requiring a high degree of trust [8]. As such, one can distinguish between two categories of ML algorithms: interpretable and non-interpretable. One example for interpretable models are Bayesian Rule Lists (BRL) [9]. By presenting itself as *if...then...else* lists, it is easy for humans to comprehend both the decision making and the individual influence of each parameter on the outcome. In contrast, the Multilayer Perceptron (MLP) model is usually more accurate, but non-interpretable, since the weights of the nodes in the hidden layers are all that is exposed to the outside. Due to the fact that different loss and activation functions take effect when updating those weights, the abstraction to the original input data is too cumbersome for a human to grasp. By the same token, in the case of ensemble approaches such as Random Forest (RF), the number of constituent trees can be very high, e.g. >100, severely harming model intelligibility, even as the accuracy is improved.

In order to overcome the tradeoff between interpretability and accuracy, we employed knowledge distillation techniques, by means of which the complex inner workings of black-box algorithms are ‘condensed’ into easy-to-understand terms. Knowledge distillation is achieved, for example, by training an interpretable model on the predictions of a more accurate, non-interpretable model, a procedure termed mimic learning [10]. By means of this technique, we are able to gain insight into the complex model’s decision process, thereby enhancing its intelligibility. As a further knowledge distillation technique, we utilized model-based feature importance for the RF model to visualize its most important features, illuminating the behavior the ‘black box’.

Our contribution consists of developing and scrutinizing a Clinical Prediction Model (CPM) to prognosticate patient-specific outcomes after hemodialysis in the Intensive Care Unit (ICU). The research set-up is modeled in Figure 1 using a Fundamental Modeling Concepts (FMC) block diagram [11]. We evaluated the performance of two different model categories, BRL and Logistic Regression (LR) as the interpretable variants, along with MLP and RF as their non-interpretable counterparts. After that, we employed mimic learning and feature importance to help overcome the tradeoff between accuracy and interpretability and provide some insight into the decision parameters of the non-interpretable algorithms. We then interviewed an expert in the field of Nephrology to scrutinize the models thus developed.

The remainder of the work is structured as follows: In Section II we place our work in the context of extant research. We present our incorporated data and models in Section III and present results of our work in Section IV. We discuss our findings in Section V followed by the conclusion in Section VII.

II. Related Work

Our work is positioned at the intersection of ML and interpretability approaches in the context of predictive modeling. For this reason, in the following, we provide an overview of existing prognostic models applied in hemodialysis using both traditional and ML-based methods. Additionally, we outline selected interpretability methods with which knowledge distillation can be achieved.

A. Predictive Models for Hemodialysis Outcomes

When it comes specifically to predictive models for hemodialysis outcomes that employ logistic or Cox regression, a clear focus on prediction of mortality for chronic hemodialysis patients can be ascertained. For instance, a predictive model developed by Marks et al. for a cohort of chronic kidney disease patients (N=3,396) presented limited results in the prediction of 5-years mortality with Area Under the Receiver Operating Characteristic Curve (AUCROC)=0.753 [12]. For 60-day mortality of maintenance hemodialysis patients, Cohen et al. achieved AUCROC=0.87, albeit in a relative small cohort of 514 patients from eight clinics [7]. Finally, a systematic literature review and external validation study conducted by Ramspek et al. indicated that AUCROC of the models validated ranged from 0.710 to 0.752 with Floege et al.’s model being the best-performing, with AUCROC=0.79 in their original population (N=11,508) [14, 13].

ML research in Nephrology has been traditionally geared towards kidney disease detection using decision trees and naïve Bayes [15, 16]. However, those models tend to be less accurate when compared to more advanced models, which prompted the community to experiment with other methods, such as Support Vector Machines (SVM) and Artificial Neural Network (ANN) for prediction of kidney disease with encouraging results [17, 18]. In a similar fashion, Lakshmi et al. compared the three models, namely, logistic regression, random forest and ANN, proposing the latter for better performance and accuracy [19].

In the specific context of hemodialysis outcomes, ML approaches have also been employed, achieving some degree of success in the chronic setting. For example, Martínez-Martínez et al. employed a range of different ML methods, such as SVM and MLP, to predict hemoglobin levels and thereby anemia in a cohort of N=13,011 patients, achieving the lowest mean absolute error (0.662) with a bagging approach [20]. Furthermore, in a comparison of three different techniques, ANN, LR and Decision Tree (DT), Srisawat et al. recommended ANN for the mortality prediction task [21].

For critically ill patients, based on a cohort of N=76 Srisawat et al. found a panel of urinary biomarkers to be strongly predictive of renal recovery, presenting an AUCROC of 0.94. Regardless of the small sample size which demands more thorough validation, the needed biomarkers are not necessarily always available in an intensive care setting, potentially limiting the applicability of this biomarker panel.

B. Knowledge Distillation

The increasing complexity of ML models and the many parameters influencing their output make it considerably difficult – if not impossible – for a human to understand the influence of any specific feature on the training and outcomes of the model. Case in point are the weights of the multiple neurons in a MLP or the potentially hundreds of trees in a RF. To enable us to ‘peek into the black box’ we employed the concept of knowlegde distillation put forth by Che et al. utilizing mimic learning [10]. In addition, algorithms such as RF make it possible to derive feature importance based on specific criteria such as mean decrease in impurity. This method provides even further insight into the algorithm’s inner workings.

In the context of ML models and results, Doshi-Velez defines interpretability as the ability to explain or to present in understandable terms to a human [22]. In contrast, Lipton

2019, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

sees interpretability as a “non-monolithic concept” which encompasses a host of “distinct ideas” [23]. Expanding on these ideas, a fledging community of researchers, deemed Fairness, Accountability, Transparency (FAT) academics, emphasizes, amongst others, explainability as one of the core principles for accountable algorithms [24]. This principle establishes that algorithmic decisions should be intelligible to end-users in “non-technical terms”. In the context of this paper, we define interpretability as a *property of machine learning algorithms and their outputs which allows scrutiny by medical experts*. Under scrutiny, we mean the ability of doctors to 1) easily ascertain the ‘reasoning’ behind an algorithm’s decision, 2) identify the most important features for the output and 3) illuminate possible biases within the model.

In effect, the enhanced performance with modern ML tools, however, is achieved at the expense of model interpretability. The ability to explain and interpret decision is a key requirement in medical applications. In the context of ML, Lipton places particular focus on identifying decision boundaries and ascertaining the influence of specific feature for improved interpretability [23]. Approaches have been developed to achieve interpretability of black-box models, such as the classification vectors approach by Baehrens et al. and the Locally-Interpretable Model-agnostic Explanations (LIME) by Ribeiro et al. [25, 26]. In particular, Katuwal and Chen applied the LIME technique for achieving interpretability of random forests for predicting ICU mortality, achieving accuracies of 80 % [8]. Still in the medical domain, Hayn et al. quantified the influence of individual features on particular decisions made by a random forest in clinical modeling applications [27].

Unlike previous work, we focus specifically on the task of outcome prediction of hemodialysis patients in intensive care while comparing two types of models side-by-side, one interpretable (BRL and LR) and another non-interpretable (MLP and RF). For aiding the interpretability of the complex models, we made use of the mimic learning technique as proposed by Che et al. in lieu of the LIME method employed in extant research, because we aim to obtain a global understanding of the model’s inner workings rather than explain individual instances of classification [8, 10]. Che et al. used Gradient Boosting Trees as mimic learning model while we applied Bayesian Ridge Regression (BRR) since their output more closely resembles logistic regression, a technique widely employed in medicine.

Given the extant literature on hemodialysis and knowledge distillation, one can ascertain a lack of works 1) using ML with a focus on critically ill patients, 2) covering different outcomes, not only mortality prediction and 3) scrutinizing model features by means of interpretability approaches. This paper addresses these research gaps.

III. Methods

In the following, we share details about the methods and data employed for the clinical models developed. We used *RapidMiner* [28], which allowed us to prepare data, develop and cross-validate first models. The final models were subsequently implemented with the *scikit-learn* library [29] in Python 2.7. The data we used were provided by the MIMIC-III dataset [30] stored in an in-memory database via an Open Database Connectivity (ODBC) interface [31]. To evaluate the models, we utilized discrimination as measured by Area Under the Receiver Operating Characteristic Curve (AUCROC) and calibration using Brier score and calibration plots. A metric routinely used in the medical context, Diagnostic Odds Ratio (DOR) was also provided in combination with precision, recall and sensitivity [32].

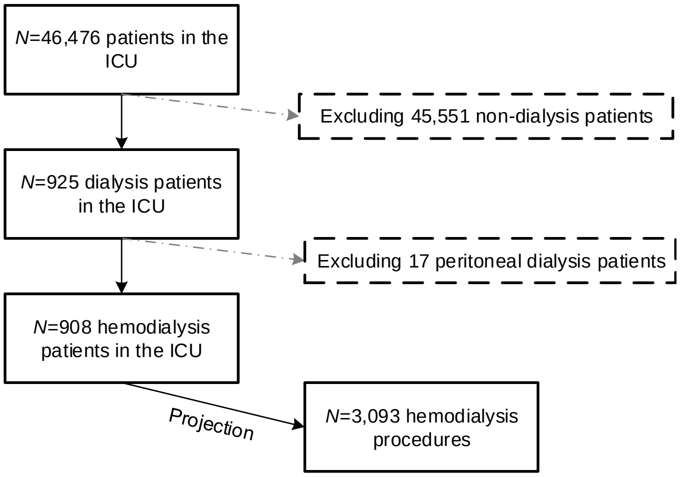

N=46,476 patients in the ICU
Excluding 45,551 non-dialysis patients
N=925 dialysis patients in the ICU
Excluding 17 peritoneal dialysis patients
N=908 hemodialysis patients in the ICU
Projection
N=3,093 hemodialysis procedures

Figure 2. Cohort selection of the hemodialysis procedures based on the MIMIC-III intensive care patients.

A. MIMIC-III Database

The MIMIC-III intensive care research database contained hospital admission data for patients collected over an eleven-year period in a Boston hospital [30]. As seen in Figure 2, out of the approximately 46,000 patients present in the dataset, we extracted 908 relevant patients for this paper, totaling approximately 3,093 hemodialysis procedures for model training. We had to exclude from the analysis patients who had undergone peritoneal dialysis, who are not relevant in an acute context.

The cohort did not contain patients who underwent hemofiltration or hemodiafiltration, only hemodialysis patients. Under hemodialysis, the data comprises both Continuous Renal Replacement Therapy (CRRT) and Intermittent Hemodialysis (IHD) modalities, therefore RRT type was a feature in the final model. As such, we derived another cohort only with CRRT patients (*N*=1,163 procedures) and IHD patients (*N*=1,930 procedures) to ascertain whether results were consistent across hemodialysis modalities. We further derived a cohort consisting exclusively of acute patients (*N*=954 procedures), since patients who developed Acute Kidney Injury (AKI) without previous history of renal disease exhibit peculiarities from a clinical standpoint.

Missing Data: Due to the manually curated nature of the MIMIC-III dataset, aside from occasional data inconsistencies, a significant amount of data was missing. For example, the columns containing serum creatinine and Glomerular Filtration Rate (GFR) values before the procedure were missing in approx. 20 % of samples. As the scikit-learn models need a complete dataset for training, we decided to impute the missing values using k-nearest neighbors algorithm (k-NN) [33].

2019, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

B. Features and Outcomes

In cooperation with a German university hospital, we conducted interviews in order to curate a list of suitable features, amounting to about 80 predictors. Those included patient demographics, such as age or Body-Mass Index (BMI), RRT parameters such as the duration of the procedure, comorbidities as well as laboratory values, including parameters such as serum creatinine and GFR for 24, 48 and 72 hours before the procedure and patient vitals.

Additionally, we included outcomes such as 90-day mortality, renal recovery, mechanical ventilation days and length of stay in the ICU. The variables ventilations days and length of stay presented continuous values, which had to be binarized for the BRL classifier to work, since it only supports binary outcomes. The complete list of features can be examined in Table A.I. The outcomes were thus defined:

- **90-days Mortality:** Indicates whether the patient has died within a 90-day period (1 = dead / 0 = alive),
- **Renal Recovery:** If patient has been for more than 7 days without hemodialysis requirement, renal function is considered to be restored (1 = recovery / 0 = no recovery),
- **Ventilation Days:** Indicates whether the patient has been on ventilation for been less than seven days (1 = true / 0 = false), and
- **Length of Stay:** Points out if length of stay has been less than 7 days (1 = true / 0 = false).

C. Modeling Algorithms

In the following, we describe the models and strategies used as well as the parameters chosen for training for both the interpretable and non-interpretable algorithms.

1) Bayesian Rule Lists: We chose the existing Python 2 implementation of BRL [9]. Letham et al. describe it as a direct competitor to decision tree approaches, as the model achieves high accuracy for classification tasks while still being intelligible for subject-matter experts. This algorithm tries to derive *if...then...else* statements over a dataset with the important criteria of their being sparse for better human readability. It builds Bayesian association rules consisting of an antecedent a and a consequent b. The consequent has a multinomial distribution over all the predicted labels y, so that the rules are defined by Equation (1):

$$a \to y \sim Multinomial(\theta) \tag{1}$$

The rules are generated by mining antecedents directly from the data and afterwards computing the posterior consequent distribution over the antecedent lists. BRL have the advantage of being easy to interpret due to their sparsity while retaining accuracy in classification. However, there are algorithms providing a higher accuracy, which also have the capability of more elaborate parameter tuning. Additionally, the current implementation of BRL has the shortcoming of a very long runtime and only being able to classify binary targets. Thus, we had to adjust the target features accordingly through use of a binary operator for continuous predictors.

Parameters: The sole adjustable parameter in the implementation used was the maximum number of iterations. Multiple adjustments to this parameter – including changes by a factor of ten – did not result in a significant change, neither for the runtime nor for the accuracy. For the evaluation, we chose a value of 50,000 maximum iterations.

2) Logistic Regression: LR is widely used for clinical prediction model development. It provides fast training time and easy-to-interpret coefficients for each model feature. For the sake of illustration, in a univariate logistic regression model, the probably that an input vector X can be assigned to the default class (or y =1, i.e., AKI onset) is given by Equation (2) also known as logit function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{2}$$

The parameters β_0 and β_1 are not known and therefore must be estimated. This algorithm seeks to derive coefficients β_i for each input feature so that they map to a binary output while minimizing the error between predicted and actual class membership using maximum-likelihood estimation [34].

Owing to its simplicity, however, LR tends to perform worse when compared to more sophisticated algorithms such as MLP or RF. Critically, LR is built upon the assumption of linearly correlated inputs and outputs. This is potentially an issue, since in a medical context one cannot necessarily assume linear relationships.

Parameters: One of the key hyperparameters to be tuned for LR refers to the regularization strength. Model performance upon validation can be improved by penalizing large coefficients, potentially reducing overfitting. As such, model sparsity is improved by a strong regularization, typically defined as λ. Another key parameter to tune is the the type of penalty for the regularization, namely L1 (lasso) or L2 (ridge). Since utilizing L1 penalty shrinks the coefficients of less importance to zero, some features might be removed altogether, a desirable property when dealing with wide datasets. For our experiment, we chose $\lambda = 1$ and L1 regularization.

3) Multilayer Perceptron: We chose the scikit-learn implementation of MLP, which is able to handle both regression and classification tasks. Just as other implementations, this network consists of multiple layers of so-called "neurons": one input layer with as many neurons as there are inputs, one output layer with the size of the number of target features and hidden layers varying in size and quantity. The log-loss function is optimized through updating weights for each neuron for each iteration of model training. The neural network can be defined as mathematical function $f(x)$ as shown in Equation (3) with the activation function K and k-times $g_i(x)$ representing the dependencies between functions with an individual weight w_i.

$$f(x) = K\left(\sum_{i=1}^{k} w_i g_i(x)\right) \tag{3}$$

MLP is a widely used algorithm in ML due to its versatility and potentially high accuracy. It provides a wealth of parameters to tune. As such, finding the right ones for a specific use case can prove cumbersome. Furthermore, the decision making process of such a neural network is not comprehensible to a human and thus provides nearly no interpretability.

Parameters: The amount of parameters to be adjusted when using neural networks is very extensive. Performing grid search over selected parameters, we found the default ones provided by the library to perform the best. This means the learning rate, which determines the speed and accuracy of convergence, was set to 0.001. The activation function, determining the output of the neurons in the hidden layer, was the rectifier linear unit "relu". The network consisted of one hidden layer with 100 neurons. We set the maximum number of iterations before convergence to 200.

4) Random Forest: The RF algorithm builds an ensemble of multiple trees in order to get a more accurate and stable prediction in comparison to an approach that relies on single decision tree. The ensemble's constituent trees utilize a random subset of the features available to split the nodes to be classified [35]. As a result of 'pooling' or majority voting of individual predictions, characteristically, RF are less prone to overfitting than regular decision trees. RF relies on bagging or bootstrap aggregation, i.e., sampling with replacement, to select samples of the training data, in an effort to reduce variance in the prediction function [36]. Hastie et al. formalize the concept in Algorithm 1.

Given a set of constituent trees b where $b \in \{1, \dots, B\}$, we denote the overall class prediction of the random forest rf over all B trees for input x by $\widehat{C}^B_{rf}(x)$. Accordingly, if we denote the class prediction of the bth constituent tree by $\widehat{C}_b(x)$, the classification output of the RF model is given by Equation (4):

$$\widehat{C}^B_{rf}(x) = majority\ vote\{\widehat{C}_b(x)\}_1^B \tag{4}$$

Algorithm 1: Training a Random Forest

```
Input: Training Data
Result: Ensemble of Trees
for $b = 1$ to $B$ do
    (a) Obtain bootstrap sample of size $N$ from training data;
    (b) Grow tree $T_b$ to the bootstrapped data, applying these steps recursively, until minimum node size $n_{min}$ is reached:
        i. Select $m$ variables at random from the available $p$ variables;
        ii. Pick the best variable/split point among $m$;
        iii. Split the node into two daughter nodes;
end
Return Ensemble of Trees $\{T_b\}_1^B$ ;
```

Parameters: The RF algorithm tends to perform well even without extensive tuning, what may explain its wide popularity [36]. In addition to the usual hyperparameters for decision trees, such as tree depth, the library employed exposes a number of hyperparameters that can be tuned specifically for RF. They include, e.g., the number of constituent trees (or estimators), i.e., B from Algorithm 1, number of variables m to split a node and the minimum number of leaves required to split an internal node. We determined the best hyperparameter combination for our use case via gridsearch, with a total number of estimators of 300, maximum tree depth of 16 and maximum number of features of eight.

D. Knowledge Distillation

In the following, the knowledge distillation approaches employed are presented in detail.

1) Mimic Learning: To provide some insight into the workings of the complex models employed we utilized a method called mimic learning. Building upon the approach of Che et al. we trained an interpretable model – the thus termed mimic model – on the outputs of the non-interpretable models, i.e., MLP and RF. In this approach, the mimic model takes on the same input features as the non-interpretable model.

In the case of MLP, the outputs of the non-interpretable model are termed soft scores. More generally, they are called prediction probabilities, meaning continuous variables approximating the actual prediction target. Training the mimic model on the prediction probabilities allows us to create a much smaller, thus understandable, faster but still comparably accurate model. In fact, under certain circumstances, it is even possible for the mimic model to generalize better than the non-interpretable model [10]. This happens because the non-interpretable model filters out certain noise in the training data, which could have a negative impact on training performance of the interpretable model. For the mimic model, we needed an algorithm which was able to predict continuous scores in order to train it on the aforementioned soft scores. For this purpose, we utilized BRR.

Similarly to common linear regression, BRR tries to find coefficients for each input feature so that they map to the target feature, minimizing loss. In addition to parameters common to linear regression, it includes regularization parameters to control the growth of the coefficients. Therefore, this model is less prone to over-fit while still being as fast as linear regression. Furthermore, regression in general has the advantage of being very fast concerning training time and interpretable, as one can easily inspect the coefficients for each feature. However, due to the simplicity of regression models, they usually lack accuracy when compared to more elaborate algorithms. Very few parameters can be adjusted for this algorithm and for our experiments, we applied the default ones. This means that all regularization parameters were set to 10^{-6} and the number of iterations before convergence was set to 300.

The process logic implemented for the mimic learning approach for MLP is shown in pseudo-code in Algorithm 2. A similar logic can be followed for RF, in which case the soft scores are replaced by prediction probabilities.

2) Feature Importance: Besides the aforementioned mimic learning approach, we provided feature importance metrics for RF the algorithm. In tree-based methods such as RF, one can estimate the relative feature importance by computing the decrease in node impurity by using it as split criterion. This decrease is averaged across all constituent trees and weighted proportionally to the number of samples it splits, i.e., nodes closer to the root of the tree will be deemed more important [37].

If we define $v(s_t)$ as the variable used in split s_t and $p(t) = N_t/N$ as the proportion of samples reaching t, the importance of a variable X_m over all N_T trees, i.e., $Imp(X_m)$, is defined by Equation (5). Note that $p(t)\Delta i(s_t, t)$ represents the weighted decrease in impurity over all nodes t which include X_m.

$$Imp(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T: v(s_t) = X_m} p(t)\Delta i(s_t, t) \qquad (5)$$

```
Algorithm 2: Mimic Learning with BRR
    Input: MLP Model, Training Dataset and Test Dataset
    Result: Sorted mimic regression coefficients
    Obtain soft scores from MPL on Training dataset;
    Train BRR model on soft scores and Training dataset;
    Apply trained BRR model on Test dataset;
    Obtain BRR regression coefficients on Test dataset;
    Sort regression coefficients;
    Return Regression Coefficients;
```

IV. Results

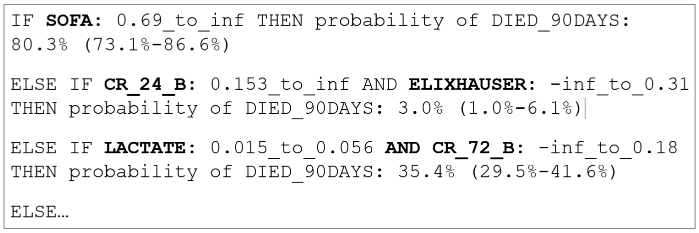
In the following section, we compare the performance of our interpretable models, BRL and LR, and our non-interpretable models, MLP and RF in terms of discrimination and calibration. Further, we present the knowledge distillation results of applying mimic learning to both MLP and RF and inspecting the feature importances of RF, since these were often the best-performing algorithms.

A. Model Performance

In the following, we assess model performance along three dimensions, discriminative power, calibration and computational performance in terms of runtimes.

1) Discrimination: Table I shows the overall performance of the employed classifiers according to the AUCROC performance metric. As expected, the MLP outperforms the BRL classifier in virtually every patient cohort and patient outcomes, excepting the prediction for ventilation days. The mimic approach using BRL trailed right along the MLP, presenting somewhat similar results. It worth noting that, in general, RF presented comparable performance to MLP, excepting the renal recovery task, in which MLP displayed more favorable results (0.91 vs. 0.83). In particular, the cohort of IHD patients presented similarly high AUCROC values for MLP, BRL, and BRR in the task of renal recovery ($\geq$ 0.9). This result suggests that patients in this cohort who do recover renal function possess very strongly discriminative features, which were captured by the algorithms.

While critically important, AUCROC is limited in the extent to which it can be used as sole metric to compare classifiers. Particularly in the medical domain, the trade-off between sensitivity (recall) and precision is highly dependent on the concrete use case. Therefore, we present further metrics in Table II for the outcome 90-days mortality in the complete patient cohort. This table shows that RF presented the best results across the metrics under analysis. Furthermore, while in terms of AUCROC MLP and RF do not differ substantially, the exception being the outcome renal recovery, there are marked differences when it comes to the other discrimination metrics, particularly DOR. LR presents overall poor results, displaying the lowest DOR. In combination its limited AUCROC, these metrics suggest that LR is likely an ill-suited choice for the task at hand in comparison with other modeling approaches. In effect, as illustrated by Table III, in comparison with previous discrimination results for similar albeit not identical tasks, our best model for renal recovery performed as well as the biomarker-based method proposed by Srisawat et al. [21].



```
IF SOFA: 0.69_to_inf THEN probability of DIED_90DAYS:
80.3% (73.1%-86.6%)

ELSE IF CR_24_B: 0.153_to_inf AND ELIXHAUSER: -inf_to_0.31
THEN probability of DIED_90DAYS: 3.0% (1.0%-6.1%)

ELSE IF LACTATE: 0.015_to_0.056 AND CR_72_B: -inf_to_0.18
THEN probability of DIED_90DAYS: 35.4% (29.5%-41.6%)

ELSE…
```

Figure 3. Excerpt of the rules from the Bayesian Rule Lists classifier when predicting 90-day mortality. Abbreviations: SOFA = Sequential Organ Failure Assessment score, CR_24_B, CR_72_B = Serum Creatinine 24h and 72h before procedure, respectively.

2) Calibration: Calibration estimates the agreement between predicted and observed risk. This is particularly relevant when it comes to predictive models employed in prognostic settings, such as ours, in which one is interested to predict future risk. In particular, it is possible for a model to be highly discriminative while over/underestimating risk, i.e., presenting poor calibration [38].

Figure 4 presents calibration curves for both MLP and RF for the outcome 90-days mortality in the complete patient cohort. In general, the MLP presents better calibration in comparison to RF. Nevertheless, we can observe that the MLP classifier tends to slightly underestimate the risk of death as the actual risk increases. In contrast, the RF classifier overestimates the probability of death in the low risk zone, with an inverse relation as the actual risk increases. We employed sigmoid and isotonic calibration to examine whether calibration could be improved. A slight improvement could be obtained for MLP, but in the case of MLP the calibration did not have the desired effect.

3) Runtimes: Concerning runtimes, there were considerable differences between the employed classifiers. While the MLP and RF took only a few seconds to conduct the full training with the configuration described previously, the BRL needed up to one hour to train on the same data. Due to the interpretable nature of the BRL, a medical expert can analyze the importance of single features directly on the model output.

B. Knowledge Distillation

Figure 3 shows the influence of some features and their values on the prediction of 90-days mortality for the complete cohort with the BRL algorithm. For this outcome, the Sequential Organ Failure Assessment (SOFA) score was a key feature. This score is widely used in intensive care for this very purpose, therefore the BRL classifier correctly detected

2019, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I. Simulation results displaying AUCROC for the different analysis cohorts and patient outcomes. Abbreviations: IHD = Intermittent Hemodialysis, CRRT = Continuous Renal Replacement Therapy, MLP = Multilayer Perceptron, RF = Random Forest, BRL = Bayesian Rule Lists, LR = Logistic Regression and BRR = Bayesian Ridge Regression, LOS ICU = Length of Stay in the ICU.

Outcome	Complete cohort					Acute patients					IHD patients					CRRT patients				
	MLP	RF	BRL	LR	BRR	MLP	RF	BRL	LR	BRR	MLP	RF	BRL	LR	BRR	MLP	RF	BRL	LR	BRR
90-days mortality	**0.84**	**0.84**	0.76	0.71	0.79	0.83	**0.85**	0.79	0.79	0.81	**0.83**	0.82	0.79	0.69	0.79	0.77	**0.78**	0.72	0.66	0.72
Renal Recovery	**0.91**	0.83	0.88	0.77	0.88	**0.86**	0.83	0.68	0.72	0.79	**0.92**	0.81	0.90	0.76	0.90	**0.86**	0.73	0.79	0.70	0.84
Ventilation Days	**0.81**	0.79	0.75	0.74	0.80	0.64	0.64	**0.68**	**0.68**	0.65	**0.81**	0.74	0.78	0.73	0.79	0.77	0.64	**0.79**	0.64	**0.79**
LOS ICU	**0.83**	0.80	0.82	0.73	0.82	**0.78**	0.64	0.69	0.63	0.73	0.80	**0.82**	0.78	0.78	0.80	0.73	0.64	**0.73**	0.63	**0.73**

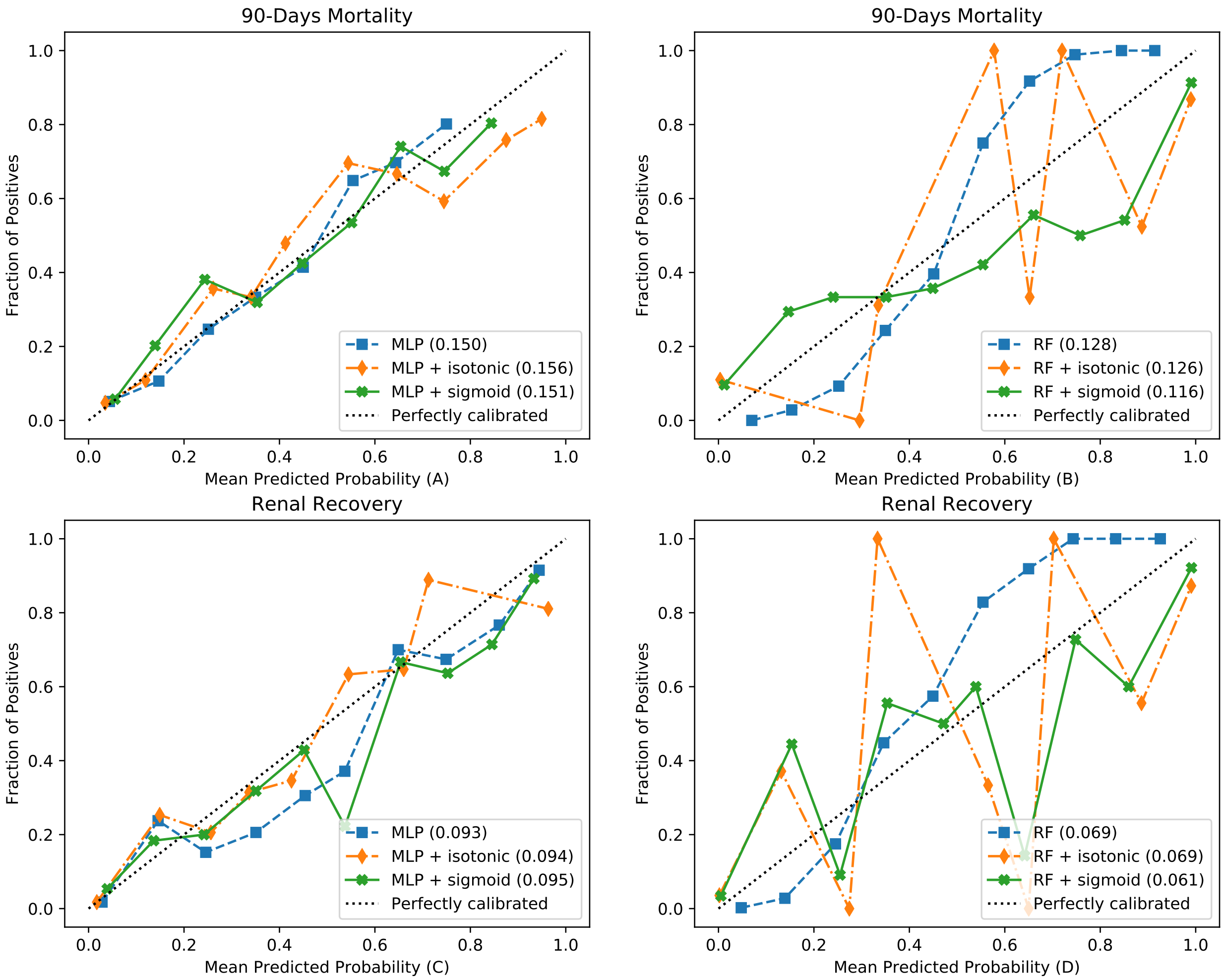


Figure 4. Model calibration depicted for Multilayer Perceptron (MLP) and Random Forest (RF) for the outcomes of 90-days mortality and renal recovery.

TABLE II. Overview of key discrimination metrics for the outcome 90-days mortality in the complete patient cohort. Abbreviations: MLP=Multilayer Perceptron, RF=Random Forest, BRL=Bayesian Rule List, LR=Logistic Regression, DOR=Diagnostic Odds Ratio.

Algorithm	Precision	Recall	Specificity	DOR
MLP	0.76	0.73	0.83	13.84
RF	0.88	0.77	0.92	41.45
BRL	0.76	0.64	0.85	10.5
LR	0.68	0.66	0.77	6.73

TABLE III. Overview of CPMs for used in the context of hemodialysis outcomes. Abbreviations: CPM=Clinical Prediction Model; N=number of patients; AUC=Area Under the Curve

CPM	*N*	End point	AUC
Marks et al.	3,396	Mortality	0.75
Cohen et al.	514	Mortality	0.87
Floege et al.	11,508	Mortality	0.79
Srisawat et al.	54	Renal recovery	0.94
Our approach	908	Renal recovery	0.92

this. “CR_24_B” corresponds to blood creatinine 24h before the hemodialysis procedure and Elixauser is a comorbidity score. High values for both of these features are associated with increased mortality, but from the output of the BRL alone it is hard to ascertain whether it correctly captured this relationship.

For the MLP and RF results to be inspected, we had to apply the mimic learning strategy discussed previously. First, we needed to evaluate if the performance of the mimic model is satisfactory when being trained on the outputs (soft scores) of the MLP. One can verify in Table I that, while the BRR is still worse than the MLP as a rule, it performed better than the BRL, even if by a small margin. It is important to highlight, however, that the mimic classifier is only as good as the predictor it originally learned from.

In Figure 5, we can assess the influence of single features on a positive prediction of both 90-day mortality and recovery of renal function. The chart depicts the regression coefficients of the BRR mimic model for MLP and RF. The sign of the coefficients determine the direction of correlation and the absolute component represents the magnitude. For example, the higher the rightmost feature, e.g., the age of the patient, the higher is the probability of the patient to die within 90 days. Conversely, the higher the features with negative coefficients, e.g., the hemoglobin value in the blood of the patient, the less likely the patient is to die within 90 days. These results were submitted to the appraisal of a Nephrology expert to establish clinical relevance and adequacy.

In addition to the mimic learning results, the RF makes it possible to derive feature importances, which enables non-ML experts to have a sense of how individual features contribute to the outcomes. Figure 6 displays feature importances for 90-days mortality in the complete cohort, reverse-ordered by feature importance. Note that we show only the top 20 features in Figure 6. Unlike the coefficients of the mimic learning approach, the information regarding the direction of correlation is not readily available, solely the magnitude of importance.

V. Discussion

In the following, we will examine the model performance in light of related work and the insights obtained via knowledge distillation.

A. Model Performance

From a classification performance standpoint, our experiments suggest MLP and RF as suitable classifiers for the given prediction tasks, with BRL as a close third. In fact, both MLP and RF performed particularly well for renal recovery prediction, a key outcome for nephrologists. In effect, the positive discriminative results obtained with these two classifiers are consistent with previous work targeted at other but similar tasks [19, 17]. When it comes to dialysis outcomes models, based on discriminative performance as measured by AUCROC, our models outperform existing work based on logistic regression approaches, except for the mortality prediction outcomes of Cohen et al. [12, 7, 14]. These works are based on the chronic setting, though. A direct comparison of the models' performance, while not advisable, gives us at least a benchmark against which to compare, since the works in the acute setting are lacking in the literature.

The only work considering a cohort of acute patients, the biomarker-based approach by Srisawat et al., performs better than our model in the task of renal recovery, but the small sample utilized in the study might compromise its generalizability [21]. Furthermore, the needed urinary biomarkers might not be readily available at all times in the ICU setting. Potential cost considerations for these biomarkers should also be taken into account.

In spite of the promising results, upon closer examination, the approaches we developed have issues that might hinder their adoption in clinical practice. If we examine, e.g., the overall best-performing classifier, RF, for the outcome 90-days mortality, it presents a higher specificity than sensitivity (recall), meaning that it will fail to acknowledge high-risk patients (true positives) more often than it identifies low-risk patients (true negatives). This behavior of the RF classifier is further illustrated by the calibration curves in Figure 4. The implications of this difference must be examined in the context of the specific clinical use case and can be mitigated with careful threshold selection and other calibration techniques.

B. Knowledge Distillation

When it comes to ML models deployed in sensitive domains, discriminative performance is not enough. The models must be scrutinized with regards to their medical relevance and physiological meaningfulness. For example, some of the features deemed important for the MLP classifier do make sense from a medical standpoint, such as higher age correlating with a higher chance of mortality. However, the results also indicate that high levels of Glomerular Filtration Rate (GFR), a measure of how well the kidneys are functioning, is associated with higher mortality, a counterintuitive outcome, since physiologically it represents a protective factor. The mimic model for the RF classifier captures similar variables as the ones found in the mimic MLP, albeit with different coefficients, i.e., with differences in magnitude. However, the same criticism can be leveled at it: GFR features prominently in it as a risk factor instead of protective factor.

In a similar fashion, for the renal recovery outcome, both mimic MLP and mimic RF captured similar features. In this case, the positive outcome (recovery) is one (1) and the negative outcomes (no recovery) is zero (0). High hemodialysis dosage, therefore, would correlate with a higher likelihood of recovery. However, there is considerable debate in the medical literature as to whether higher dosage leads to better outcomes [12]. Therefore, this result must be interpreted with caution. Furthermore, the Sequential Organ Failure Assessment (SOFA) score appears to be a factor *favoring* renal recovery in the case of mimic RF, what clearly contradicts clinical expectations. A hypothetical explanation for this scenario is that patients with high SOFA scores are particularly ill and therefore receive, on average, better standard of care than other less severely ill patients. This hypothesis suggests that the prediction results should be further stratified by disease severity.

It is important to note that these potential spurious correlations are only illuminated through model interpretability, be it because of the nature of the model or the application of mimic learning. Thus, the model interpretability approach employed gives us the possibility to examine the correlations and question assumptions which otherwise might just go unnoticed when using non-interpretable models. However, usually there are non-linear correlations between certain blood values and outcome (e.g., U-shaped curve), such as potassium, as either too low or too large values can influence the patient's

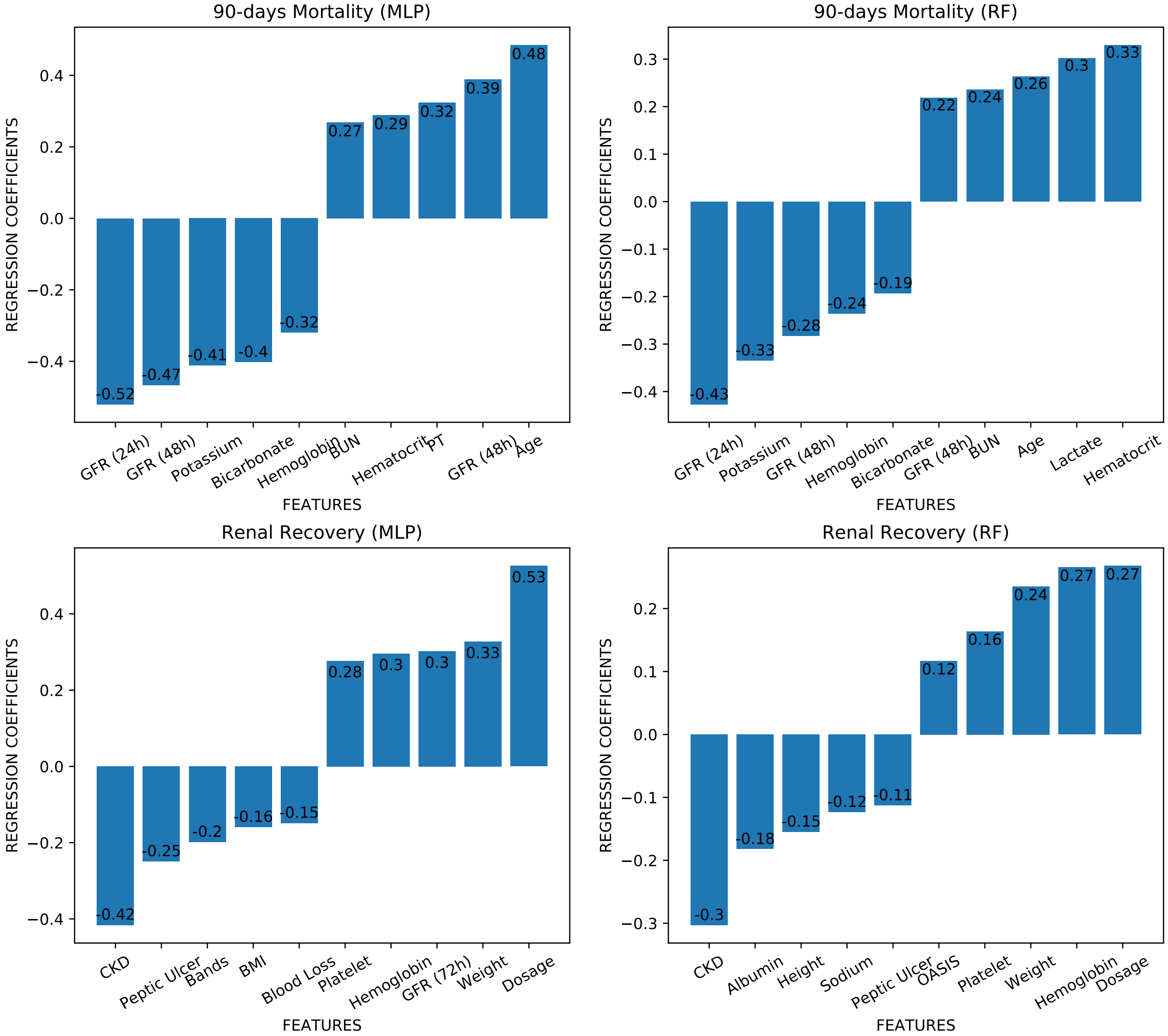


Figure 5. Coefficients of the most important features for the Bayesian Ridge Regression trained as mimic model for 90-days mortality and renal recovery for both Multilayer Perceptron (MLP) and Random Forest (RF). Abbreviations: GFR = Glomerular Filtration Rate 24h and 48h before procedure, respectively; BUN = Blood Urea Nitrogen; CKD = Chronic Kidney Disease; OASIS = Oxford Severity of Illness Score.

health negatively. Such relationships cannot be adequately represented by the mimic learning approach utilized.

To a certain extent, the feature importances of the RF classifier reflected the knowledge gleaned from applying the mimic learning approach, highlighting some of the parameters that also were captured in the mimic learning method, such as hematocrits, lactate, blood urea nitrogen and glomerular filtration rate. Despite this fact, when examining all approaches in combination, there is disagreement, for instance, in the magnitude of contribution or in how often a feature is mentioned across different techniques. Hall and Gill recommend that researchers combine different interpretability approaches in order to obtain a more intelligible picture of the model's behavior [39].

Finally, algorithms considered to be interpretable might not necessarily be intelligible. This is particularly evident for the BRL algorithm. Take its output as depicted in Figure 3. As a matter of fact, higher lactate values usually lead to other complications, but the upper bound of "infinity" is not meaningful in clinical practice. In order to refine and validate those assumptions, it is necessary to further analyze the data. Finding actual upper and lower bounds in the dataset can provide some insight into the actual values the model considers when making predictions.

VI. Limitations

Even though we achieved satisfactory discriminative performance, this analysis was based on a comparatively small patient cohort (N=908). Therefore, a validation study with a larger cohort is needed in order to derive generalizable claims. Additionally, missing data may have a significant influence on the quality of the predictions and certain features could be dropped if they are missing a large amount of values. We sought to mitigate this effect by means of multiple imputation with k-NN, but we cannot guarantee that no biases resulted from this approach.

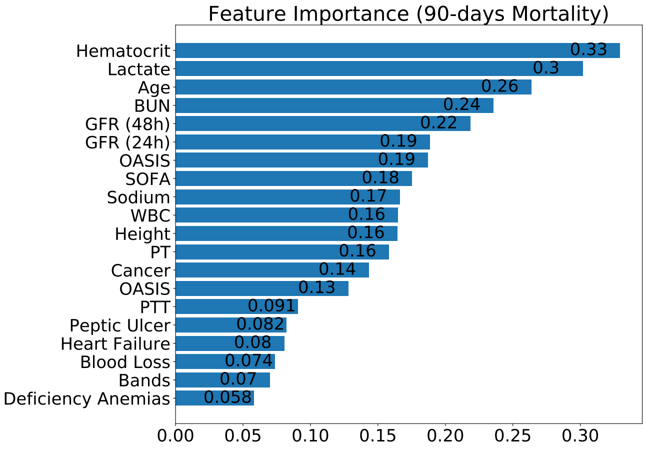


Figure 6. Feature importances for 90-days mortality based on the Random Forest (RF) classifier. Abbreviations: BUN = Blood Urea Nitrogen; GFR = Glomerular Filtration Rate 24h and 48h before procedure, respectively; OASIS = Oxford Severity of Illness Score; SOFA = Sequential Organ Failure Assessment; WBC = Whole Blood Count; PTT = Partial Thromboplastin Time.

Furthermore, in most cases, the mimic learning approach is fundamentally limited by the performance of the original model. In our experiments, the BRR performed worse when being trained on the outputs of the MLP as opposed to being trained directly on the real targets, because it most probably also assimilates the errors of the MLP. This can be ameliorated by improving the performance of the MLP through further parameter tuning.

Besides, the literature of interpretability approaches is growing. In this paper, we were able only to utilize two of them, mimic learning and method-based feature importances, thus necessarily providing a limited picture of model behavior. Other methods could be explored, for example involving local interpretability, such as Local Interpretable Model-agnostic Explanations (LIME) [26]. Finally, the medical relevance and physiological meaningfulness of the mimic models was evaluated by one expert only. Ideally, these should be assessed by a panel of expert to reduce biases.

VII. Conclusion

In this paper, we compared the performance of different models used in the prediction of hemodialysis outcomes, namely, 90-days mortality, ventilation days, length of ICU stay and renal recovery using data routinely acquired in a intensive care setting. The algorithms employed consisted of a combination interpretable and non-interpretable models. Our results suggest that ML approaches such as MLP and RF present satisfactory discriminative results (AUCROC $\leq$ 10 in the case of renal recovery) when compared with interpretable algorithms, such as LR or BRL.

However, an important aspect is the interpretability of such models if they are to be used for decision support in a clinical setting. To this end, we applied a knowledge distillation technique called mimic learning along with feature importances in order to scrutinize the 'black-box' best-performing models. The use of these techniques made it possible to uncover potentially spurious correlations captured by the algorithms, therefore shedding light on model biases. Therefore, we urge researchers who rely on ML for clinical predictive modeling to include an assessment of possible biases using for example knowledge distillation approaches.

Future work could take the form of further data analysis and processing, i.e., inclusion of more features, more elaborate imputation strategy and collection of more information about the patients. Besides, deployment in a clinical setting requires external validation using datasets from different institutions. Subsequently, an impact analysis of the use of such models in a clinical setting should be conducted to ascertain the impacts on care delivery and patient outcomes.

Acknowledgments

Parts of the given work were generously supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 780495.

References

[1] H. F. da Cruz, S. Horschig, and M. Schapranow, "Prediction of Patient Outcomes After Renal Replacement Therapy in Intensive Care," in *Proceedings of the 3rd Intl. Conf. on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing (HEALTHINFO)*, vol. 3, 2018, pp. 7–12.

[2] E. Marieb and S. M. Keller, *Essentials of Human Anatomy & Physiology*. Pearson, 2018.

[3] C. Ronco *et al.*, "Renal Replacement Therapy in Acute Kidney Injury: Controversy and Consensus," *Critical Care*, vol. 19, no. 1, pp. 1–11, 2015.

[4] G. M. Fleming, "Renal Replacement Therapy Review," *Organogenesis*, vol. 7, no. 1, pp. 2–12, 2011.

[5] S. Negi *et al.*, "Renal Replacement Therapy for Acute Kidney Injury," *Renal Replacement Therapy*, vol. 2, no. 1, p. 31, 2016.

[6] L. Forni and P. Hilton, "Continuous Hemofiltration in the Treatment of Acute Renal Failure," *New England Journal of Medicine*, vol. 336, no. 18, pp. 1303–1309, 1997.

[7] L. M. Cohen *et al.*, "Predicting Six-month Mortality for Patients Who Are on Maintenance Hemodialysis," *Clin J Am Soc Nephrol*, vol. 5, no. 1, pp. 72–79, jan 2010.

[8] G. J. Katuwal and R. Chen, "Machine Learning Model Interpretability for Precision Medicine," https://arxiv.org/abs/1610.09045 [retrieved: August, 2018], Oct 2016.

[9] B. Letham *et al.*, "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model," *Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[10] Z. Che *et al.*, "Interpretable Deep Models for ICU Outcome Prediction," in *Proceedings of the AMIA Annual Symposium*, vol. 2, 2016, pp. 371–380.

[11] A. Knöpfel, B. Gröne, and P. Tabeling, "Fundamental modeling concepts," *Effective Communication of IT Systems, England*, p. 51, 2005.

[12] A. Marks *et al.*, "Predicting Renal Replacement Outcomes in a Large Community Cohort with Chronic Kidney Disease," *Nephrol Dial Transplant*, vol. 30, no. May, pp. 1507–1517, 2015.

[13] C. L. Ramspek *et al.*, "Prediction Models for the Mortality Risk in Chronic Dialysis Patients: a Systematic Review and Independent External Validation Study," *Clinical Epidemiology*, vol. 9, pp. 451–464, 2017.

[14] J. Floege *et al.*, "Development and Validation of a Predictive Mortality Risk Score from a European Hemodialysis Cohort," *Kidney International*, vol. 87, no. 5, pp. 996–1008, 2015.

2019, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

[15] P. S. Baby and P. Vital, "Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms," *J of Eng Res & Tech*, vol. 4, no. 07, pp. 206–210, 2015.

[16] R. Greco *et al.*, "Decisional Trees in Renal Transplant Follow-up," *Transplant Proc*, vol. 42, no. 4, pp. 1134–1136, may 2010.

[17] S. Vijayarani and S. Dhayanand, "Kidney Disease Prediction Using SVM and ANN," *Comp Bus Res*, vol. 6, no. 2, pp. 6–17, 2015.

[18] P. Sinha and P. Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," *J of Eng Res & Tech*, vol. 4, no. 12, pp. 608–612, 2015.

[19] K. R. Lakshmi, Y. Nagesh, and M. Veerakrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability," *J of Adv in Eng & Tech*, vol. 7, no. 1, pp. 242–254, 2014.

[20] J. M. Martínez-Martínez *et al.*, "Prediction of the Hemoglobin Level in Hemodialysis Patients using Machine Learning Techniques," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 2, pp. 208–217, 2014.

[21] N. Srisawat *et al.*, "Urinary Biomarkers and Renal Recovery in Critically," *Clin J Am Soc Nephrol*, vol. 6, pp. 1815–1823, 2011.

[22] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," no. Ml, pp. 1–13, 2017.

[23] Z. C. Lipton, "The Mythos of Model Interpretability," in *Proceedings of the Workshop on Human Interpretability in Machine Learning*, 2016, pp. 96–100.

[24] D. Shin and Y. J. Park, "Role of fairness, accountability, and transparency in algorithmic affordance," *Computers in Human Behavior*, vol. 98, pp. 277–284, 2019.

[25] D. Baehrens *et al.*, "How to Explain Individual Classification Decisions," *Mach Learning Res*, vol. 11, pp. 1803–1831, 2010.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proceedings of the Workshop on Human Interpretability in Machine Learning*, 2016, pp. 91–95.

[27] D. Hayn *et al.*, "Plausibility of Individual Decisions from Random Forests in Clinical Predictive Modelling Applications," *Studies in Health Technoly and Informatics*, pp. 328–335, 2017.

[28] M. Hofmann and R. Klinkenberg, *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press, 2013.

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J of Machine Learning Res*, vol. 12, pp. 2825–2830, 2011.

[30] A. E. W. Johnson *et al.*, "MIMIC-III, a Freely Accessible Critical Care Database," *Scientific Data*, vol. 3, 2016.

[31] F. Färber *et al.*, "SAP HANA database," *ACM SIGMOD Record*, vol. 40, no. 4, pp. 45–51, 2012.

[32] A. S. Glas *et al.*, "The Diagnostic Odds Ratio: A Single Indicator of Test Performance," *Journal of Clinical Epidemiology*, vol. 56, no. 11, pp. 1129–1135, 2003.

[33] J. M. Keller, M. R. Gray, and J. A. Givens, "A Fuzzy K-nearest Neighbor Algorithm," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 4, pp. 580–585, jul 1985.

[34] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013, vol. 398.

[35] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 1st ed. O'Reilly Media, Inc., 2017.

[36] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009, vol. 1.

[37] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding Variable Importances in Forests of Randomized Trees," *Neural Information Processing Systems*, pp. 1–9, 2013.

[38] L. E. Hodgson *et al.*, "Systematic Review of Prognostic Prediction Models for Acute Kidney Injury (AKI) in General Hospital Populations," *BMJ Open*, vol. 7, no. 9, pp. 1–10, 2017.

[39] P. Hall and N. Gill, *An Introduction to Machine Learning Interpretability*, 2018.

Appendix

In Table A.I, we share the complete list of features used for our models.

TABLE A.I. Model features. Note that related features are grouped together. Abbreviations: Body-Mass Index (BMI), Acute Kidney Injury (AKI), Sequential Organ Failure Assessment (SOFA), Simplified Acute Physiology Score (SAPS), Partial Thromboplastin Time (PTT), International Normalized Ratio (INR), Prothrombin Time (PT), Whole Blood Count (WBC).

Category	Feature
Demographics	Age
	Height, Weight, BMI
	Ethnicity
	Gender
Hemodialysis-related	Dosage
	Modality
	AKI stage
Comorbidities	AIDS
	Alcohol abuse
	Blood loss anemia
	Cardiac arrhythmias
	Chronic pulmonary
	Coagulopathy
	Congestive heart failure
	Deficiency anemias
	Depression
	Diabetes complicated, Diabetes uncomplicated
	Drug abuse
	Elixhauser Vanwalraven score
	Fluid electrolyte imbalance
	Hypertension
	Hypothyroidism
	Liver disease
	Lymphoma, Metastatic cancer, Solid tumor
	Obesity
	Other neurological disorders
	Paralysis
	Peptic ulcer
	Peripheral vascular
	Psychoses
	Pulmonary circulation
	Renal failure
	Rheumatoid arthritis
	Valvular disease
	Weight loss
ICU scores	OASIS
	SOFA
	SOFA Renal
	SAPS
Vitals	Heart rate
	Systolic Blood pressure
	Diastolic Blood pressure
	Mean Blood pressure
	Respiratory Rate
	Temperature °C
	Oxigen Saturation (SpO_2)
Laboratory values	Anion gap
	Albumin
	Bands
	Bicarbonate
	Bilirubin
	Blood urea nitrogen
	Creatinine 24, 48 and 72h before procedure
	Chloride
	Glucose
	Hematocrit
	Hemoglobin
	Lactate
	Platelet
	Potassium
	PTT, INR, PT
	Sodium
	WBC
	Glomerular Filtration Rate 24, 48 and 72h before procedure

2019, © Copyright by authors, Published under agreement with IARIA - www.iaria.org