Prediction of Patient Outcomes after Renal Replacement Therapy in Intensive Care

Harry F. da Cruz*, Siegfried Horschig*, Christian Nusshag**, and Matthieu-P. Schapranow*
* Hasso Plattner Institute, Digital Health Center, Rudolf-Breitscheid-Str. 187, 14482 Potsdam, Germany
** Heidelberg University Hospital, Department of Nephrology, Im Neuenheimer Feld 162, 69120 Heidelberg, Germany

E-Mail: *{harry.freitasdacruz|schapranow}@hpi.de, siegfried.horschig@student.hpi.de,
**christian.nusshag@med.uni-heidelberg.de

***Abstract*—In order to compensate severe impairments of renal function, artificial, extracorporeal devices have been developed to enable Renal Replacement Therapy. The parameters utilized for this procedure and the specific patient characteristics substantially affect individual patient outcomes and overall disease courses. In this paper, we present a clinical prediction model for outcomes of critically ill patients who underwent a specific form of renal replacement, hemodialysis. For this purpose, we employed two machine-learning models: Bayesian Rule Lists and Multi-Layer Perceptron. To provide more transparency to the perceptron model, we applied mimic learning to its output based on a Bayesian Ridge Regression model. Results show that while the perceptron model outperforms the rule-based classifier, the use of the mimic learning approach enables more thorough model scrutiny by a medical expert, revealing possible model biases, which might have gone unnoticed, a sensitive issue in a high-stakes domain such as medicine.**

***Keywords*–Clinical Prediction Model; Renal Replacement Therapy; Machine Learning; Supervised Learning.**

I. INTRODUCTION

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 homoeostasis [1]. If this system is impaired severely, e.g., in the context of an Acute Kidney Injury (AKI), artificial, extracorporeal organ replacement therapy becomes necessary [2]. Therefore, different Renal Replacement Therapy (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 [3].

Dialysis outcomes are highly dependent on both the patient's characteristics and clinical parameters, as well as on the type of the RRT procedure applied [4]. Furthermore, RRT modalities based on a filtration circuit, such as hemofiltration or hemodiafiltration are particularly costly, requiring specialized equipment and nursing staff [5]. 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 [6].

Aside from usual criteria like accuracy or recall, when employing a machine learning model in the medical context, one especially important factor is the interpretability of the model, since doctors must take full responsibility for the respective decision and therefore require a high degree of trust on the model [7]. As such, one can roughly distinguish between two categories of machine learning algorithms: interpretable and non-interpretable. One example for interpretable models are Bayesian Rule Lists (BRL) [8]. 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 Multi-Layer 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 just too cumbersome for a human to grasp.

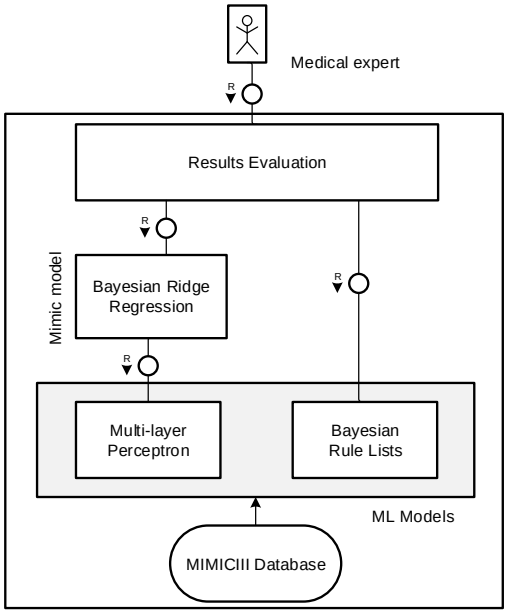

Figure 1. Our machine-learning setup modeled as a FMC block diagram. We incorporated a multi-layer perceptron and bayesian rule lists machine-learning model.

In order to overcome the tradeoff between interpretability and accuracy, we employed a strategy called mimic learning [9]. By training an interpretable model on the predictions of the more accurate, non-interpretable model, we gain insight into its decision process and can therefore enhance the non-interpretable model's intelligibility.

Our contribution consists of a Clinical Prediction Model (CPM) to prognosticate patient-specific outcomes after RRT in the Intensive Care Unit (ICU) as modeled in Figure 1 using Fundamental Modeling Concepts (FMC) block diagram. We evaluated the performance of two different models, BRL as the interpretable variant and MLP as its non-interpretable counterpart. After that, we employed mimic learning to help

 ISBN: 978-1-61208-675-0

overcome the tradeoff between accuracy and interpretability and provide some insight into the decision parameters of the MLP. We then interviewed an expert in the field of Nephrology for scrutiny of the models thus developed.

The remainder of the work is structured as follows: In Section II we place our work in the context of related work. 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 VI.

II. Related Work

Machine learning research in Nephrology has been traditionally geared towards kidney disease detection using decision trees and naïve Bayes [10, 11]. However, those models tend to be less accurate when compared to more advanced models, which prompted the community to experiment with other methods. Vijayarani and Dhayanand and Sinha and Sinha used Support Vector Machine (SVM) and Artificial Neural Network (ANN) for prediction of kidney disease with encouraging results [12, 13]. In a similar fashion, Lakshmi et al. compared the three models regression, random forest and ANN, proposing the latter for better performance and accuracy [14].

The enhanced performance with modern machine learning 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 machine learning, Lipton placed the particular focus was on identifying decision boundaries and ascertaining the influence of specific feature for improved interpretability [15]. 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. [16, 17]. In particular, Katuwal and Chen applied the LIME technique for achieving interpretability of random forests for predicting ICU mortality, achieving accuracies of 80 % [7]. 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 [18].

Unlike previous work, we focus specifically on the task of outcome prediction of RRT patients while comparing two types of models side-by-side, one interpretable (BRL) and another non-interpretable (MLP). For aiding the interpretability of the complex model, 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 [7, 9]. 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.

III. Methods

In the following, we share details about methods and data employed for our clinical models.

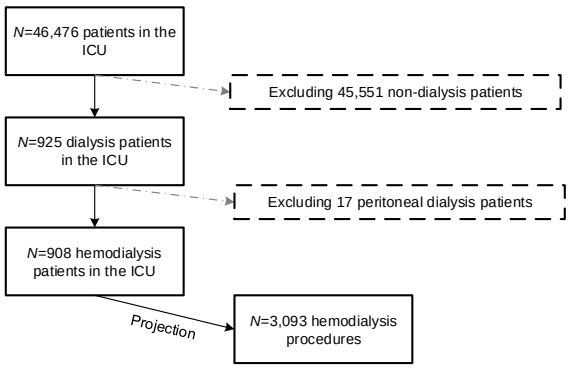

Figure 2. Cohort selection of the dialysis procedures based on the MIMIC III intensive care patients.

A. Tools

We used *RapidMiner* [19], which allowed us to prepare data, develop and cross-validate first models. The final models were subsequently implemented with the *scikit-learn* library [20] in Python 2.7. The data we used were provided by the MIMIC III dataset [21] stored in an in-memory database via an Open Database Connectivity interface [22].

B. Data

The *MIMIC III* dataset contained hospital admission data for patient collected over an eleven-year period in a Boston hospital. 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 3,000 dialysis procedures for model training. We had to exclude from the analysis patients who had undergone peritoneal dialysis, since they are not relevant in an acute context.

The cohort does 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. We therefore 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 dialysis modalities. We further derived a cohort consisting exclusively of acute patients (N=954 procedures) since patients, who presented acute kidney injury without previous history of renal disease, present peculiarities from a clinical standpoint.

In cooperation with the *Nierenzentrum Heidelberg*, we conducted interviews with a subject-matter expert 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 lab values, including parameters such as serum creatinine and Glomerular Filtration Rate (GFR) for 24, 48 and 72 hours before the procedure. Additionally, we included patient vitals and outcomes such as 90-day mortality, renal recovery, mechanical ventilation days and length of stay in the ICU.

 ISBN: 978-1-61208-675-0

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 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) [23].

C. Models

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, as well as the interpretability approach employed.

1) Interpretable - Bayesian Rule Lists: For the interpretable model, we chose the existing Python 2 implementation of BRL [8]. 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 in Equation 1.

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

Mining antecedents from the data generated these rules 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 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 – incl. 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) Non-Interpretable: Deep Neural Network: As non-interpretable model, 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 2 with the activation function K and k-times $g_i(x)$ representing the dependencies between functions with an individual weight w_i.

$$f(x) = K\left(\sum_{i=1}^{k} w_i g_i(x)\right) \tag{2}$$

MLP are a widely used form of machine learning due to their versatility and high accuracy. They provide a wealth of parameters to tune, but 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 high. Performing grid search over some of the parameters, we found the default ones from 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 was set to 200.

3) Interpretability Approach: Mimic Learning: The large amount of neurons in the MLP and the many parameters influencing their weights and output make it very difficult – if not impossible – for a human to understand the influence of each feature on the training. Therefore, we aimed to provide some insight into the workings of the MLP by applying a method called mimic learning. Building upon the approach of Che et al. we trained an interpretable model – the so-called mimic model – on the outputs of the non-interpretable model (MLP). In this approach, the mimic model takes on the same input features as the non-interpretable model.

For classification tasks, the outputs of the non-interpretable model are termed soft scores, because as they are probabilities, they are continuous variables, coming close to the actual target features. Training the mimic model on the soft scores allows creating a much smaller, thus understandable, faster but still equally accurate model. Using the principle Che et al. called knowledge distillation, it is even possible for the mimic model to generalize better than the non-interpretable model [9]. 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 Bayesian Ridge Regression.

Bayesian Ridge Regression: Similar to common linear regression, this algorithm tries to find coefficients for each input feature so that they map to the target feature, minimizing loss. In addition to common 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 were set to 300.

 ISBN: 978-1-61208-675-0

The process logic implemented for the mimic learning approach is shown in pseudo-code in Algorithm 1.

Algorithm 1: 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 model, the BRL, and our non-interpretable model, the MLP. Although there were continuous values for our target variables in the dataset, we had to transform them into a binary format in order for the BRL classifier to work. Therefore, we considered the following outcomes:

- **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 dialysis 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).

The complete list of features can be found in Table A.I.

Table I shows general performance of the employed classifiers according to the AUCROC performance metric. As expected, the MLP outperforms the BRL classifier in for virtually every patient cohort and patient outcomes, excepting the prediction for ventilation days. The mimic approach using BRL trailed right along the MLP, presenting similar results. Concerning runtimes, there were considerable differences between the two classifiers. While the MLP 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.

Figure 3 shows the influence of some features and their values on the prediction of 90-day mortality. 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 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.

```
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.

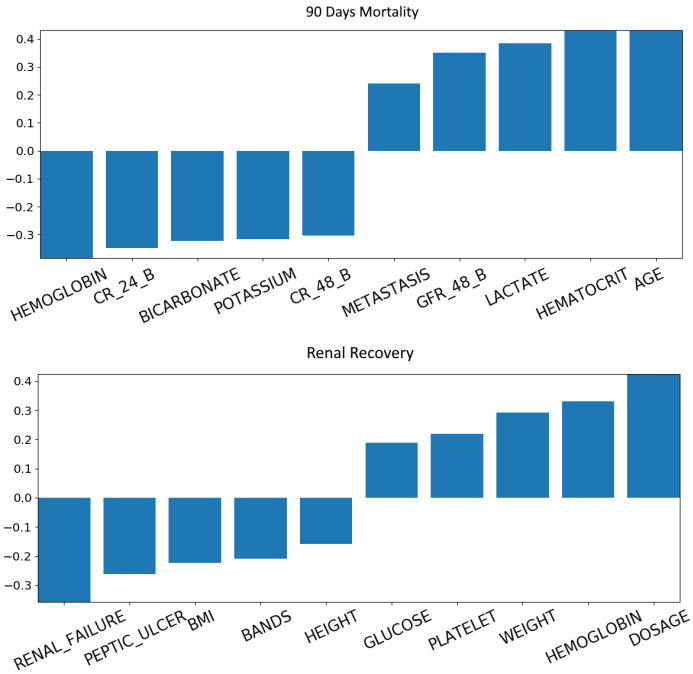

Figure 4. Coefficients of the most important features (last five and top five after sorting) for the Bayesian Ridge Regression trained as mimic model for 90-days mortality and renal recovery. Abbreviations: CR_24_B, CR_48_B, GFR_48_B = Serum Creatinine and Glomerular Filtration Rate 24h and 48h before procedure, respectively, BMI = Body-Mass Index.

For the MLP results to be inspected, we had to apply the mimic learning strategy discussed. 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 that, while the BRR is still worse than the MLP, it performed better than the BRL, if only 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 4, we can assess the influence of single features on a positive prediction of both 90-day mortality and recovery of renal function. 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 leftmost feature, 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.

 ISBN: 978-1-61208-675-0

Outcome	Complete cohort			Acute patients			IHD patients			CRRT patients		
	MLP	BRL	BRR	MLP	BRL	BRR	MLP	BRL	BRR	MLP	BRL	BRR
90-days mortality	0.84	0.76	0.79	0.83	0.79	0.81	0.83	0.74	0.79	0.77	0.72	0.72
Recovery of renal function	0.91	0.88	0.88	0.86	0.68	0.79	0.90	0.87	0.91	0.86	0.79	0.84
Ventilation days <7	0.81	0.75	0.80	0.64	0.68	0.65	0.81	0.78	0.79	0.77	0.79	0.79
ICU stay days <7	0.83	0.82	0.82	0.78	0.69	0.73	0.80	0.78	0.80	0.73	0.73	0.73

TABLE I. Simulation results displaying Area Under the Receiver Operating Characteristic Curve (AUCROC) for the different analysis cohorts and patient outcomes. Abbreviations: IHD = Intermittent Hemodialysis, CRRT = Continuous Renal Replacement Therapy, MLP = Multi-Layer Perceptron, BRL = Bayesian Rule Lists and BRR = Bayesian Ridge Regression.

V. Evaluation and Discussion

From a classification performance standpoint, our performed experiments suggest MLP as a suitable classifier for the given tasks, with BRL as a close second. MLP performed particularly well for renal recovery prediction, a key outcome for nephrologists. However, both approaches have issues that may hinder adoption in clinical practice.

For example, some of the features deemed important for MLP 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 creatinine are associated with lower mortality, which contradict observations in clinical practice. Additionally, as per Figure 4 Glomerular Filtration Rate (GFR), a measure of how well the kidneys are functioning, is associated with higher mortality, a likewise counterintuitive outcome.

Similarly for renal recovery prediction, where high weight and glucose levels are associated with poor outcomes, what contradicts expert knowledge. Furthermore, usually there are non-linear correlations between certain blood values and mortality (e.g., U-shaped curve), such as potassium, as either too low or too large values can influence the patient's health negatively. Such relationships cannot be adequately represented by the mimic learning approach utilized.

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 create assumptions which otherwise might just go unnoticed when using non-interpretable models. The same observations apply for the output of the BRL. For instance, 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 go further with the data analysis. Finding actual upper and lower bounds in the dataset can provide some insight to the actual values the model considers when making predictions.

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. By training the regression as a mimic model, we can make assumptions on how the MLP *may* make its decisions. There still is a gap between the performance of the regression model and the MLP, which makes it difficult to say how close those coefficients are to the actual influence of features in the MLP. The mimic model performs worse when being trained on the outputs of the MLP as opposed to being trained on the real targets, because it most probably also learns the errors of the MLP. This can be a resolvable issue by improving the performance of the MLP through further parameter tuning and data preparation.

VI. Conclusion

In this paper, we compared the performance of different models when being used in the prediction in the renal context. An important part is the interpretability of such models to validate their applicability for decision support. We used a mimic learning approach to make a MLP interpretable and compared this output to that of the interpretable model, the BRL. Preliminary results for prediction of 90-day mortality enable the exploration of interpretability, showing the influence of single features.

Future work includes more elaborate use of the data, meaning inclusion of more features, more elaborate imputation strategy and collection of more information about the patients. The binary classification limitation of the current implementation of BRL can be overcome by using a more advanced algorithm, such as the one proposed by Yang et al. [24]. When a higher precision is achieved, interpretable models can be used in the context of a clinical decision support system, allowing the doctors to validate the decisions and giving patients insight into their treatment. Likewise, 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 a grant of the German Federal Ministry of Economic Affairs and Energy (01MD15005C).

 ISBN: 978-1-61208-675-0

REFERENCES

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

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

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

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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

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

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

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

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

[23] 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.

[24] H. Yang, C. Rudin, and M. Seltzer, "Scalable Bayesian Rule Lists," https://arxiv.org/pdf/1602.08610.pdf [retrieved: August, 2018], Feb. 2016.

APPENDIX

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

Category	Feature
Demographics	Age
	Height, Weight, BMI
	Ethnicity
	Gender
Dialysis-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
	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	Heartrate
	Systolic Blood pressure
	Diastolic Blood pressure
	Mean Blood pressure
	Respiratory Rate
	Temperature °C
	Oxigen Saturation (SpO_2)
Laboratory values	Aniongap
	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

TABLE A.I. Model features. Note that related features are grouped together.