Exploring ExpertRuleFit and Time-Series Models: A Comparative Study for AKI Prediction in the ICU

MLRH Group 4: Chi Him Ng, Dheeraj Varghese, and Danila Rusinkiewicz

Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Abstract. AKI, characterized by sudden kidney function loss, significantly impacts ICU patients, thereby causing prolonged stays and increased mortality. This study predicts AKI onset in the ICU using AmsterdamUMCdb data, employing ExpertRuleFit for interpretability and trustworthiness, and Time-Series models (Transformers, LSTMs) for their ability to exploit temporal patterns. It was found that current implementations of ExpertRuleFit models possess strong discrimination abilities but lack the accuracy required for reliable clinical predictions. Transformer models were effective in short-term forecasting, but encountered challenges over extended time windows. The metrics of LSTMs were erroneously inflated due to a biased sampling method, ultimately revealing their poor performance.

Keywords: AKI · ExpertRuleFit · Time-Series · AI for Health

1 Business Understanding

1.1 Acute Kidney Injury in the ICU

Acute kidney injury (AKI) encompasses a spectrum of clinical presentations of sudden excretory kidney function loss. The kidneys, composed of nephrons, play a crucial role in regulating fluid balance, electrolytes, osmolality, and pH, as well as eliminating waste and releasing hormones. However, a progressive decline in nephrons harms kidney function. AKI can disrupt homeostasis, potentially progressing to irreversible nephron loss and Chronic Kidney Disease (CKD). Therefore, AKI may lead to fluid imbalance, causing issues like hyperkalemia and abnormal sodium levels. Kidney failure negatively impacts organ systems, including the lungs and the cardiovascular system (16). AKI outcomes are influenced by the triggering disease, severity, duration, and baseline health. Presentations vary from mild to severe impairments, potentially resulting in permanent renal function loss. In high-resource settings, AKI affects one in five hospitalized adults, which is approximately half of the patients in the intensive care unit (ICU) (14). The occurrence of AKI is linked to unfavorable consequences, including prolonged ICU and hospital stays, increased risk of CKD, and higher short- and long-term mortality (13). No scientifically proven effective drug therapy exists for AKI. Therefore, treatment relies on supportive care, primarily Renal

Replacement Therapy (RRT) for severe cases in the ICU. Despite RRT and other advances in treatment, mortality rates for severe AKI patients requiring RRT persistently remain high at 50% with limited improvement (21).

Diagnosistic criteria The field of AKI research faced challenges due to inconsistent definitions, until the Acute Dialysis Quality Initiative introduced the RIFLE criteria in 2004 that defined AKI based on changes in serum creatinine from baseline and urine output specifically (6). The Kidney Disease: Improving Global Outcomes (KDIGO) AKI classification system is a widely used adaptation of RIFLE that aids in diagnosing and managing AKI. It identifies minimal changes in renal function parameters and includes three severity levels that are associated with incremental risks of adverse outcomes (see Table 1) (15).

Stage	Serum Creatinine	Urine Output
1	1.5 to 1.9 times baseline *or* $\geq$ 0.3 mg/dl ($\geq$ 26.5 μmol/l) increase	0.5 ml/kg/hour for 6 to 12 hours
2	2.0 to 2.9 times baseline	0.5 ml/kg/hour for $\geq$ 12 hours
3	3.0 times baseline *or* increase in serum creatinine to $\geq$4.0 mg/dl ($\geq$353.6 μmol/l) *or* initiation of renal replacement therapy *or* in patients <18 years a decrease in eGFR to <35 ml/minute per 1.73 m^2	<0.3 ml/kg/hour for $\geq$24 hours *or* anuria for $\geq$ 12 hours

Table 1: Staging of acute kidney injury according to KDIGO

AKI risk factors Numerous studies have explored the epidemiology of AKI, but a common challenge is the limited understanding of its pathogenesis in many cases (26). Nonetheless, established risk factors known to heighten susceptibility to AKI in hospitalized patients, as indicated by KDIGO AKI clinical practice guidelines, include Age over 65, diabetes mellitus, chronic diseases (heart, lung, liver), and anemia. Additionally, exposure to contributing factors like sepsis (especially in severe stages like septic shock), major surgery, or the use of nephrotoxins is associated with increased AKI risks (15).

1.2 Prediction of AKI with Machine Learning and Reasoning

Given the severe consequences and limited treatment options for inpatient AKI, there is growing interest in preventive tools. Artificial intelligence, particularly machine learning, holds promise in forecasting AKI onset, potentially improving patient outcomes. Previous research has prioritized supervised classification tasks in this domain. For instance, Zimmerman et al. used logistic regression and random forest models on physiological data from the MIMIC-III database to predict AKI in adult ICU patients (34). Another study employed a Gradient Boosting Machine algorithm to predict stage-2 inpatient AKI within a 12-hour window (17). Dong et al. developed an ensemble model for pediatric ICU patients, predicting AKI up to 48 hours earlier than conventional guidelines, and designed an actionable alert system for interventions based on patient context (7). In contrast, Du et al. used a Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM) to predict Sepsis-associated AKI persistence by leveraging the time-series dimension of ICU data (12).

Similar to the aforementioned studies, the current research centers on utilizing supervised machine learning techniques for AKI prediction in the ICU. What sets this paper apart is the incorporation of expert knowledge into the machine learning model using ExpertRuleFit, potentially enhancing both its explainability and trustworthiness for clinicians. Moreover, Transformer models are recognized for their ability to extract semantic correlations in long sequences (33). Consequently, a Time-Series Transformer model and a LSTM will be employed as an additional approach to assess the effectiveness of a time-series component in the predictions.

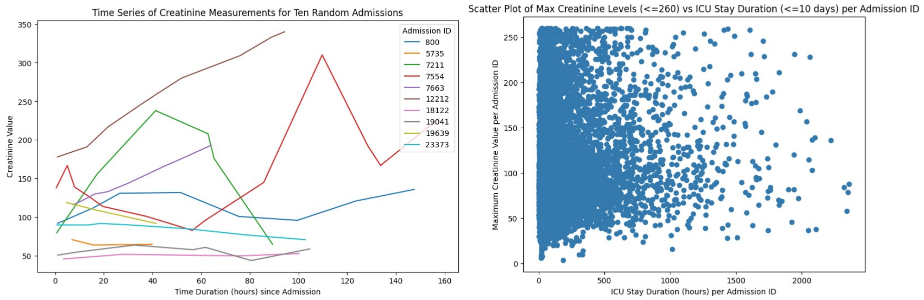
2 Data Understanding

This study utilizes clinical ICU data sourced from the AmsterdamUMCdb. This database comprises around 1 billion data points from 23,106 admissions involving 20,109 patients aged 18 and older who were admitted between 2003 and 2016. It includes anonymized information such as limited patient and admission details, patient measurements, laboratory measurements, and drug and treatment data (28). To understand what variables are relevant for the current AKI prediction task, we had to investigate both the literature and data availability in AmsterdamUMCdb.

The prominent variables in this study are creatinine and urine output, which are both integral to KDIGO diagnosis criteria. However, urine output exhibits variability influenced by non-renal factors like fluid intake. Therefore, creatinine levels are chosen to construct the AKI outcome variable. To examine the available creatinine data, a time series for 10 admissions and maximum creatinine levels were plotted. In Figure 1a, creatinine measurements, like other variables in the database, show an inconsistent measure rate, suggesting that aggregation is required for consistent time intervals. Figure 1b reveals considerable variability in maximum creatinine levels, with no discernible trend between stay duration and creatinine levels. Due to the lack of baseline creatinine data, an approximation must be constructed from creatinine measurements for each admission.

Measurement	**Count**	**Mean**	**Min**	**Max**
Creatinine	20,2700	117.53	0	3187
Heart Rate(20)	2,218,626	87.46423	-23,421	96,160
Systolic Blood pressure (Hypotension) (23)	2,109,758	131.4245	-32,697	336,170
Hemoglobin (Anemia)(15)	223,630	6.54	0	502
Glucose (Hyperglycemia)(11)	820898	8.33	-1	444,444
Calcium (27)	145,940	2.08	0	233

Table 2: Relevant AKI measurements together with statistical details



(a) Creatinine timeline for 10 admissions (b) Creatinine levels versus ICU stay

With the use of scientific literature, several other relevant variables were identified. Firstly, the aforementioned risk factors for AKI, such as advanced age, female gender and sepsis patients are considered relevant variables. Furthermore, several other variables were found to correlate with AKI in the medical literature or were utilized in AKI prediction research. These include vital signs and lab measurements, of which a subset can be found in Table 2. Here, again we observe that variables are not measured at the same frequency, with some variables containing a lot more values. Furthermore, most variables also exhibit unrealistic min and max values (i.e. Glucose at a level of 444,444), most likely representing entry errors. This must be taken into account in the feature prepossessing stage to ensure outliers will not hamper model performance. During the exploration stage, we also investigated the data availability of nephrotoxic drugs (i.e. NSAIDs) and vassopressors due to their impact on the development of AKI. We found there is a total of 1117 specific drug types administered to patients. This means that drug variables perhaps first need to be grouped to avoid unnecessary model complexity. Finally, it was found that for some relevant features, there is no or limited data available. Due to the data's anonymous nature, age, height and weight are grouped as ordinal variables (i.e. age 60-69). Furthermore, the database contains limited information on patient diagnoses (i.e. diabetes) information.

3 Data Preparation

After querying and inspecting the data in the previous section, the next step was to preprocess the data such that the predictor variables and AKI outcome variable become machine-processable. The selected features are therefore subjected to the data preparation methods described in the current section.

3.1 Outlier Detection

Autoencoders were employed for outlier detection of numerical data because of their ability to compress input data into a reduced latent space and reconstruct the original data through decoding (30; 19). Training the autoencoder to minimize the difference between input and reconstructed data enables it to capture and replicate typical data patterns (19), making it an ideal candidate for anomaly detection. For normalization, z-score normalization was applied to standardize selected columns in the datasets, ensuring uniform scaling (22). Mean Squared Error and Adam optimizer were employed during training. Outliers were identified through reconstruction errors, measuring disparities between input and reconstructed data (4). We employed Isolation Forest to detect outliers based on reconstruction errors. This method is suitable for high-dimensional data by constructing binary trees and measuring partitions to separate outliers (32). Autoencoders and isolation forests combined can learn a condensed data representation and effectively identifying outliers (2). For outlier replacement, the Gaussian distribution method was used (24).Constraints were applied to the replacement values data consistency. Depending on the data, they are adjusted to be non-negative (e.g. heart rate), ensured by taking the maximum between the computed mean and standard deviation and zero and limited based on a percentile threshold, such as the 5th percentile, of the non-outliers' data distribution. Figure 2 shows the distribution before and after outlier replacement for the urine variable. As can be seen, the distribution is almost the same while the extreme upper outlier is replaced.

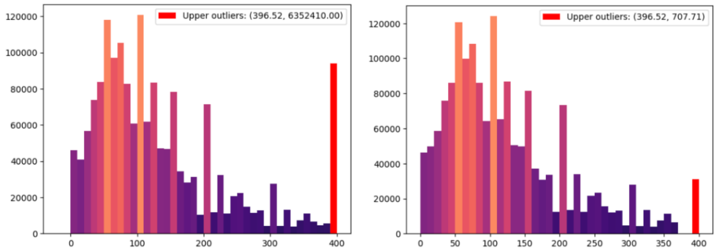


Fig. 2: Urine values before and after outlier replacement

3.2 Data Aggregation and Imputation

We performed aggregation to organize the data in 1-, 3-, 6- and 12-hour intervals for each admission. We initially converted the *measuredat* column, which indicates measurement times in milliseconds, into discrete hour intervals to group measurements (i.e. lab or vital measurements) taken simultaneously. The process involves managing two sets of data iteratively— the aggregated data and the data awaiting merging. Datasets marked for merging are grouped by *admissionid* and *hour*, employing three calculation methods: maximum (creatinine), sum aggregation (urine output), and mean (e.g., heart rate). These distinct aggregations yield sets capturing values for each admission ID across different hour intervals. The aggregated datasets were then merged based on *admissionid* and *hour*, consolidating hourly measurements for admissions. Post-merging steps involve refining the dataset. It was sorted chronologically, and the index was reset for better organization.

To address missing values, forward-filling was employed to maintain continuity, particularly since not every variable has new measurements for every time interval. For ordinal admission variables (age, weight, and height group) a K-Nearest

Neighbors imputer with a parameter of two neighbors was utilized. This imputation method fills missing values by examining the nearest data points and selecting the most common category among them. The choice of two neighbors ensures localized pattern analysis within the dataset, promoting computational efficiency, particularly in datasets with sparse clusters or localized variations (5; 25).

3.3 Feature engineering

In the domain of feature engineering, binary variables were created to indicate the intake of nephrotoxic drugs (NSAIDs, ACEIs, ARBs) or vasopressors. Utilizing the "Farmacotherapeutisch Kompas" [1], measurements of individual drugs were grouped to create these features. Another binary feature denoting whether a patient received mechanical ventilation at a specific time was introduced, and the admission type variable was leveraged to generate features representing various surgery types, such as "Cardiac Surgery". To introduce a temporal dimension in supervised learning, temporal features were integrated. These features capture changes in vital signs or lab values compared to the previous time step (i.e. heart rate or glucose change). Additionally, a variable indicating the number of days the patient had spent in the ICU up to that specific point in time was included.

For the AKI outcome Label, a binary label was derived from the creatinine level measurements. Here, '1' denotes the patient having AKI (at least stage 1 according to KDIGO diagnostic criterion) at that point in time. The lowest creatinine measurement in the first 24 hours of admission was considered the baseline creatinine level for each admission. The AKI outcome label was computed for each creatinine measurement during an admission.

3.4 Exlusion criteria and final data

This study excluded patients with ICU stays under 48 hours to ensure dataset stability when predicting outcomes within a 12-hour timeframe. Shorter ICU stays may lack sufficient data points for accurate AKI assessment. Additionally, patients without baseline creatinine measurements or levels exceeding 260 mmol/liter were excluded, as predicting AKI requires comparing changes from this baseline. Higher baseline levels may hinder accurate AKI identification.

After the aforementioned feature preprocessing steps, datasets for four different time windows (1, 3, 6, and 12 hours) were obtained, containing variables outlined in Table 3 (excluding the AKI outcome variable). The datasets underwent an approximate 80-20 train-test split, with PatientID separation to prevent data leakage.

4 Modeling

For the modeling task, two specific (supervised) machine learning models will be utilized: ExpertRuleFit and a time-series transformer model.

[1] Farmacotherapeutisch Kompas

Patient info	Gender, weightgroup, height group, surgery type (cardiac, vascular, trauma, gastroenterology, lungs, oncology, neuro), shock, sepsis
Vital Signs	Urine output, heart rate, temperature, blood pressure, respiratory rate, oxygen saturation
Lab Values	Serum creatinine, hemoglobine, glucose, calcium, kalium, pH, thrombocytes, sodium, billubirine, hematocryt, lactate, leukocyten
Treatment	Mechanical ventilation, vassopressors, NSAIDs, ARBs, ACEIs
Temporal	Changes in measurements compared to previous timestep, length of ICU stay (in days)

Table 3: Final set of features

4.1 ExpertRuleFit

The ExpertRuleFit method (ERF), a derivative of the RuleFit model, is a rule-based ML ensemble capable of learning sparse linear models with automatically detected interaction effects in the form of decision rules. Its interpretability aligns with traditional linear models. Leveraging its rule-learning capability, ExpertRuleFit is well-suited for large datasets, facilitating automatic knowledge discovery. By integrating expert knowledge in the form of rules and linear terms, ExpertRuleFit complements RuleFit, incorporating exceptional cases and correcting spurious patterns. This integration enhances human trust in model results, especially in medical contexts, where physicians' deeper understanding of human physiology and symptoms is crucial for diagnosis. ExpertRuleFit employs the standard Lasso method for regularization, applying tailored penalties to weight coefficients and eliminating uninformative rules and linear terms. Different Lasso penalties can be used for the data-derived, confirmed or optional expert rules, contributing to the refinement of the model's final rule set (8).

AKI Expert Rules In utilizing ExpertRuleFit for AKI prediction, rules need to be extracted from guidelines and integrated into the framework. Expert rules are assigned one of either level of evidence: confirmed or optional. For confirmed rules, we use extracted rules from the aforementioned KDIGO AKI clinical guidelines. KDIGO mentions that patients should be stratified for risk of AKI according to their susceptibilities and exposures. The rules are confirmed because a moderate level of evidence in combination with a recommendation for clinicians to follow this guideline (1B) was explicitly mentioned (15). A total of 23 confirmed rules were added. Optional rules were extracted from National Insitute of Health and Care AKI guidelines (1), medical research papers on AKI, or were created combinations of guideline rules, summing to a total of 21 optional rules and terms. Table 4 shows a set of example rules that were implemented.

Evidence	Rule	Meaning	Source
Confirmed	has_sepsis	Sepsis patient	KDIGO(15)
	cardiac_surgery	Cardiac surgery patient	
	hema < 13.5 & gender_Man	Anemia in male patient	
Optional	weightgroup in c('8')	Overweight (110+ kg)	Lan et al.(18)
	glucose >11.1 & hema <13.5	Hyperglaecemia & anemia	Gorelik et al.(11)
	systolic_ABP <90	Hypotension	NICE(1)

Table 4: Example of extracted AKI expert rules

Data Handling ERF required adjustments due to limitations in handling temporal dependencies in time sequences. Specifically, it treated each instance as a separate 'patient', disregarding temporal context. Consequently, this led to significant outcome class imbalances due to varying patient stay durations. To address this, strategies were implemented. Patients with ICU stays exceeding 5 days without developing AKI were excluded, focusing on relevant instances. We also experimented with class weights and undertook undersampling on the training datasets to manage outcome class imbalances. Here, limiting the maximum imbalance to 12 times, and using weights 0.85/0.15 for positive/negative yielded optimal test results. Finally, the dataset for the one-hour timestamp was not utilized due to significant class imbalance, which would have resulted in considerable data loss through undersampling.

Experiment The study aimed to predict AKI development at 3-, 6-, and 12-hour intervals using four ERF model variations. These were compared against a conventional RuleFit model serving as the baseline. Firstly, we implemented **Standard ERF**, this model incorporates data rules and expert knowledge with full penalties on optional expert insights ($\nu = \eta = 1$). Secondly, **ERF Optional**, this model is similar to Standard ERF but prioritizes optional expert knowledge by setting penalties at $\nu = \eta = 0.5$, following Ebnar et al.'s study without detailed justification (3). Furthermore, **ERF Only**, which integrates solely expert knowledge, excluding data rules. Lastly, **RuleFit Model**, which implements data rules exclusively without expert knowledge. Additionally, the existing **PRE** RuleFit method served as a baseline. Evaluation metrics included AUC, $f1$ score, and recall. Model complexity was gauged using the ensemble's size, and the proportion of expert knowledge was vital for interpretability. 10-fold cross-validation ensured balanced metric measurements. This method splits data into ten parts for iterative training and testing, providing comprehensive performance assessment by averaging results across iterations (9).

4.2 Time Series Approach: Transformer and LSTMs

Transformers (29) employ the attention mechanism for processing sequential data, distinguishing them from recurrent neural networks like LSTMs (10). They parallelize computation, effectively handling long-range dependencies. When applied to time series data, they face challenges such as the absence of temporal ordering and high input dimensionality. Time-series transformers (31) address these issues by modifications such as using positional embeddings to encode temporal information, and using a context window to sample the input sequence.

We employ the time-series transformer implementation from Hugging Face[2], featuring a vanilla encoder-decoder model with self-attention and feed-forward layers. Comparative performance analysis is conducted against LSTM models. Our hypothesis posits that the time-series transformer would consider periodic variations and subtle changes indicative of impending acute kidney injuries.

[2] HuggingFace Time Series Transformer Documentation

Data Handling The initial step involved encoding all categorical features into numerical values. For the transformers, each patient was considered throughout their stay. However, we require that there are sufficient values to cover both context and prediction lengths. For instance, in a 3-hour aggregated window, a context length of 32 (equivalent to 180 hours) and a prediction length of 6 (36 hours) were employed. LSTMs also considered a patient throughout their stay, employing a threshold-based sampling method to address class imbalance. For 1 and 3-hour windows, predictions were made for 30 hours at a time (30-hour lookbacks), while for 6 and 12-hour windows, this was extended beyond 30 hours to provide ample context.

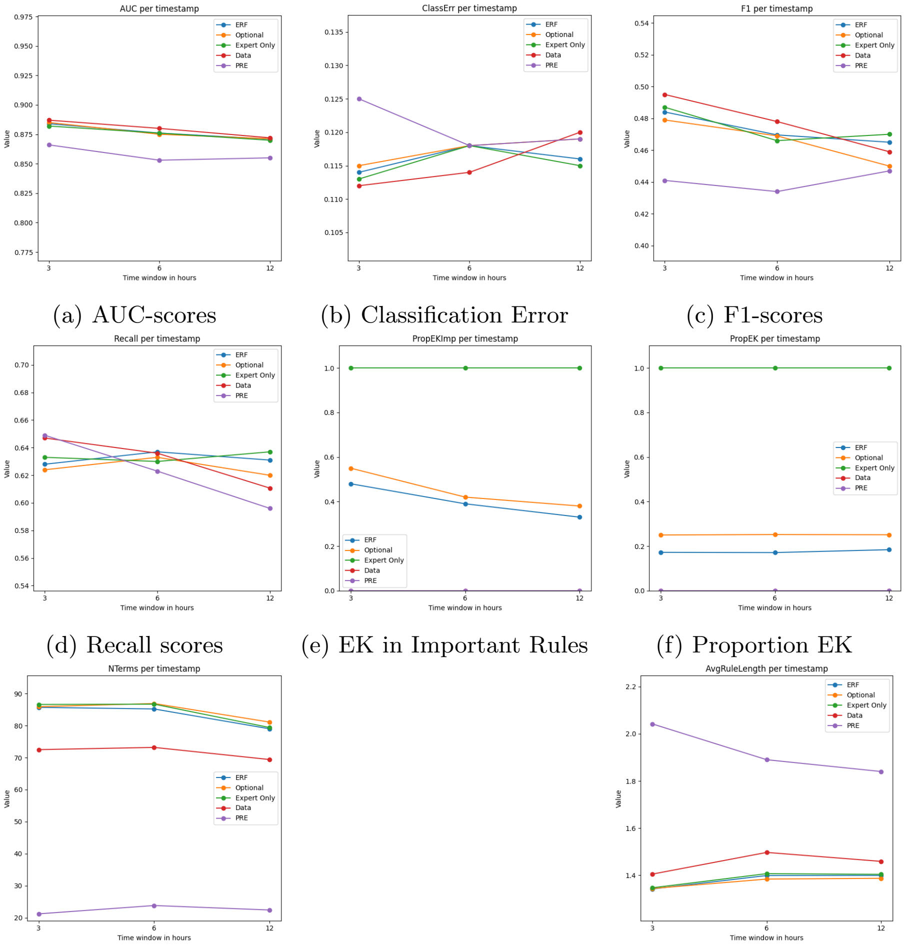
Experiment The objective of the transformer model was autoregressive generation of a binary outcome, indicating the presence of AKI. Labels were pre-propagated for 1, 3, 6, and 12-hour prediction windows. The architecture remained consistent across windows, featuring 4 encoder and decoder layers, 2 encoder attention heads, and 4 decoder attention heads. The dimensions of the attention and feed-forward layers was set to 128, resulting in approximately 1.8 million parameters. Each window-specific model underwent training for 10 epochs, utilizing a learning rate of 1e-3, AdamW optimizer, and Negative log likelihood (NLL) loss. The LSTM (and a bidirectional LSTM) was also trained for binary outcome prediction. These models comprised 4 layers with a hidden size of 50, resulting in around 80 thousand parameters. These models were trained for 30 epochs with early stopping. Model evaluations were performed using AUC and $f1$ score metrics.

5 Evaluation

This section aims to analyze and deliberate on the outcomes derived from employing the ERF method in conjunction and the time series approach.

5.1 ERF

Results The results of the ERF experiment reveal critical insights into the model's performance across different metrics. The AUC, averaging around 0.88, indicates strong overall discrimination, with the baseline PRE model showing the least effectiveness among the assessed models. However, the ERF models show similar performance to RuleFit (data), with no significant improvement. Although marginal differences exist among timestamps, the 3-hour interval displays slightly superior performance (refer to Figure 3a). Despite similar classification error rates across timestamps, the 3-hour interval exhibits slightly higher errors. However, there is a critical inadequacy in accurately classifying positive instances, evident from the F1-score and recall. The observed recall of approximately 0.63 indicates a significant proportion of misclassified positive instances, highlighting a sensitivity deficiency (see Figures 3b, 3c, and 3d). The (E)RF model comprises around 80 terms, with approximately one-fifth attributed to EK (refer to Figures 3e and 3f). ERF and RF models significantly increase term inclusion, around fourfold compared to the PRE model. Despite this increase, the average rule length remains shorter (figures 3g and 3h).



(a) AUC-scores

(b) Classification Error

(c) F1-scores

(d) Recall scores

(e) EK in Important Rules

(f) Proportion EK

(g) Number of Terms Included

(h) Length of Rules

Discussion Comparative analysis reveals consistent superior performance of ERF and its variations in terms of F1-scores, recall, AUC, and classification error. However, a challenge persists due to significant class imbalance within outcomes. Efforts to alleviate this through undersampling and weight adjustments show improvements but leave room for enhancement, particularly in balancing recall and F1-scores. In medical contexts like AKI, prioritizing recall remains crucial for accurate positive case identification. Yet, this emphasis shouldn't excessively compromise F1-scores. Furthermore, overfitting poses a risk, leading to contradictory rules, potentially capturing noise rather than genuine data patterns. Further research is needed to investigate observed discrepancies, such as negative coefficients for expert terms representing the older age groups. However, ERF results also reveal the learning of more sensible rules, as for example an increased risk of AKI was identified when *temperature*> 38.371 or *creatinine*> 190.5.

Regarding model complexity, RF and ERF incorporate a more extensive number of terms compared to PRE, whereas PRE exhibits longer rules. Longer rules in PRE may signify comprehensive decision pathways, while more terms in RF and ERF suggest a broader set of decision criteria, allowing a nuanced understanding of data patterns. However, heightened complexity may challenge interpretability and computational resources, impacting deployability and scalability.

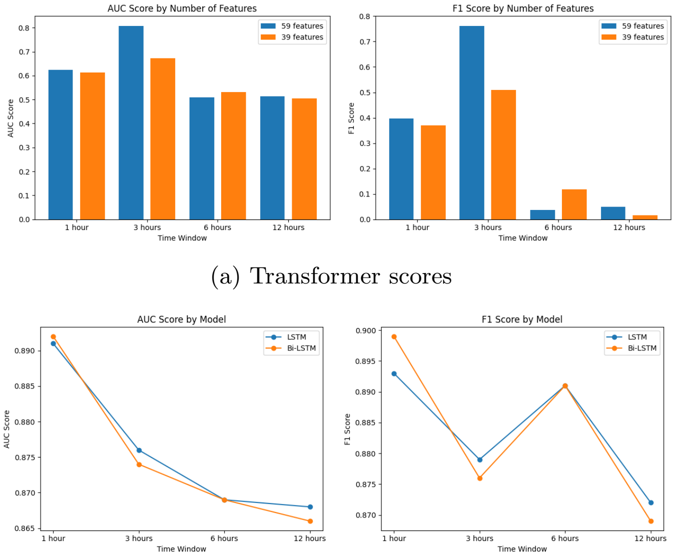
Time constraints limited extensive experimentation, particularly in addressing the class imbalance issue. Exploring alternative techniques, such as different data aggregation methods, weight adjustments or modifications to implement stratified folds, could offer more comprehensive analyses for handling class imbalances and improving performance. Overall, despite ERF models showing promising performance and interpretability, the current versions falls short of clinical usability due to its lack of sufficient accuracy. Further improvements are necessary to enhance its precision and reliability before considering its practical application in clinical contexts.

5.2 Transformer and LSTMs

Results The time-series models performance presents intriguing insights. The transformer performs best with a 3-hour time window when assessed on AUC and F1 Score (refer Figure 4a). On the 3-hour time window, an AUC of 0.808 indicates that it can discriminate well between positive and negative cases of AKI. However, the F1 Score of .762 suggests that there is still room for improvement in terms of precision and recall. The performance decreases as the time window increases to 6 and 12 hours. On the 12-hour time window, it achieves an AUC of 0.504, which indicates that it is almost random in its predictions. The F1 Score of .017 also confirms that the model has very poor performance in this setting. The effect on performance when reducing features from 59 to 39 is negligible.

Both LSTM and Bi-LSTM models have comparable performance (refer Figure 4b). The AUC and F1 Score metrics are close in values across different time windows and lookback periods. However, Bi-LSTM has a slightly higher F1 Score in the 1-hour time window. The highest AUC and F1 Score values are obtained with a 1-hour time window and a 30-hour lookback period, which suggests that the models are sensitive to recent changes in the patient's condition.

Discussion For time-series transformers, predicting AKI 6 hours or more in advance poses challenges, potentially due to data limitations or task complexity. Despite this, the Transformer excels in forecasting approximately 36 hours ahead, underscoring its ability to capture temporal nuances. Analyzing the impact of feature reduction reveals that it is less reliant on additional engineered features



(a) Transformer scores

(b) LSTMs scores

for temporal change information, showcasing a robust understanding of the temporal dimension.

In contrast, the initially reported elevated AUC and F1 Scores of the LSTM models were subsequently invalidated due to interference from threshold-based sampling in test. Upon examination, a substantial portion of the lookback segments revealed either complete absence or full presence of AKI, revealing a flaw in data selection. Conducting tests with the proper set, showed that the models had an average AUC score of 0.5 and a much lower F1.

5.3 Comparative Evaluation

In conclusion, our study aimed to predict the onset of AKI in ICU patients using clinical measurement data sourced from the AmsterdamUMCdb. Employing two distinct approaches, we utilized the binary classification method ExpertRuleFit, enriched with expert knowledge for enhanced interpretability and reliability, alongside Time-Series models (including Time-Series Transformers and LSTMs) that exploit temporal patterns inherent in the data. Our findings reveal that, in terms of clinical applicability, the current ExpertRuleFit models currently lack the necessary accuracy for producing reliable estimates, despite the addition of clinical expert knowledge. Conversely, the Time-Series models, particularly transformers, exhibit notable strengths in sensitivity to time windows, robustness to number of features, and effective forecasting, positioning it as a more suitable choice for AKI prediction. They are better than LSTMs for short-term forecasting of AKI, which can help clinical decision making. However, they face challenges for long-term forecasting, which require more data and more sophisticated models.

Bibliography

[1] Acute kidney injury: prevention, detection and management (2019), `https://www.nice.org.uk/guidance/ng148`

[2] Almansoori, M., Telek, M.: Anomaly detection using combination of autoencoder and isolation forest. In: 1st Workshop on Intelligent Infocommunication Networks, Systems and Services (WI2NS2). pp. 25–30. Budapest University of Technology and Economics (2023)

[3] Augustin, T.: Expert rulefit: Complementing rule ensembles with expert knowledge

[4] Beggel, L., Pfeiffer, M., Bischl, B.: Robust anomaly detection in images using adversarial autoencoders. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I. pp. 206–222. Springer (2020)

[5] Beretta, L., Santaniello, A.: Nearest neighbor imputation algorithms: a critical evaluation. BMC medical informatics and decision making **16**(3), 197–208 (2016)

[6] Case, J., Khan, S., Khalid, R., Khan, A., et al.: Epidemiology of acute kidney injury in the intensive care unit. Critical care research and practice **2013** (2013)

[7] Dong, J., Feng, T., Thapa-Chhetry, B., Cho, B.G., Shum, T., Inwald, D.P., Newth, C.J., Vaidya, V.U.: Machine learning model for early prediction of acute kidney injury (aki) in pediatric critical care. Critical Care **25**(1), 288 (2021)

[8] Ebner, L., Master, N., ten Teije, A., van Harmelen, F., Augustin, T.: Expert rulefit: Complementing rule ensembles with expert knowledge. KR4HC2021 (2021)

[9] Fushiki, T.: Estimation of prediction error by using k-fold cross-validation. Statistics and Computing **21**, 137–146 (2011)

[10] Gers, F., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm. Neural computation **12**, 2451–71 (10 2000). https://doi.org/10.1162/089976600300015015

[11] Gorelik, Y., Bloch-Isenberg, N., Hashoul, S., Heyman, S.N., Khamaisi, M.: Hyperglycemia on admission predicts acute kidney failure and renal functional recovery among inpatients. Journal of Clinical Medicine **11**(1), 54 (2021)

[12] He, J., Lin, J., Duan, M.: Application of machine learning to predict acute kidney disease in patients with sepsis associated acute kidney injury. Frontiers in Medicine **8**, 792974 (2021)

[13] Hoste, E.A., Bagshaw, S.M., Bellomo, R., Cely, C.M., Colman, R., Cruz, D.N., Edipidis, K., Forni, L.G., Gomersall, C.D., Govil, D., et al.: Epidemiology of acute kidney injury in critically ill patients: the multinational aki-epi study. Intensive care medicine **41**, 1411–1423 (2015)

[14] Hoste, E.A., Kellum, J.A., Selby, N.M., Zarbock, A., Palevsky, P.M., Bagshaw, S.M., Goldstein, S.L., Cerdá, J., Chawla, L.S.: Global epidemiology and outcomes of acute kidney injury. Nature Reviews Nephrology **14**(10), 607–625 (2018)

[15] Kellum, J.A., Lameire, N., Group, K.A.G.W.: Diagnosis, evaluation, and management of acute kidney injury: a kdigo summary (part 1). Critical care **17**, 1–15 (2013)

[16] Kellum, J.A., Romagnani, P., Ashuntantang, G., Ronco, C., Zarbock, A., Anders, H.J.: Acute kidney injury. Nature reviews Disease primers **7**(1), 52 (2021)

[17] Koyner, J.L., Carey, K.A., Edelson, D.P., Churpek, M.M.: The development of a machine learning inpatient acute kidney injury prediction model. Critical care medicine **46**(7), 1070–1077 (2018)

[18] Lan, J., Xu, G., Zhu, Y., Lin, C., Yan, Z., Shao, S.: Association of body mass index and acute kidney injury incidence and outcome: A systematic review and meta-analysis. Journal of Renal Nutrition (2023)

[19] Marimont, S.N., Tarroni, G.: Anomaly detection through latent space restoration using vector quantized variational autoencoders. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1764–1767. IEEE (2021)

[20] Mohamadlou, H., Lynn-Palevsky, A., Barton, C., Chettipally, U., Shieh, L., Calvert, J., Saber, N.R., Das, R.: Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. Canadian journal of kidney health and disease **5**, 2054358118776326 (2018)

[21] Negi, S., Koreeda, D., Kobayashi, S., Iwashita, Y., Shigematu, T.: Renal replacement therapy for acute kidney injury. Renal replacement therapy **2**, 1–7 (2016)

[22] Patro, S., Sahu, K.K.: Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462 (2015)

[23] Sato, R., Luthe, S.K., Nasu, M.: Blood pressure and acute kidney injury. Critical Care **21**(1), 28 (2017)

[24] Shaikh, S.A., Kitagawa, H.: Efficient distance-based outlier detection on uncertain datasets of gaussian distribution. World Wide Web **17**, 511–538 (2014)

[25] Shao, X., Wu, S., Feng, X., Song, R.: Categorical missing data imputation approach via sparse representation. International Journal of Services Technology and Management **22**(3-5), 256–270 (2016)

[26] Singbartl, K., Kellum, J.A.: Aki in the icu: definition, epidemiology, risk stratification, and outcomes. Kidney international **81**(9), 819–825 (2012)

[27] Singh, N.P., Panwar, V., Aggarwal, N.P., Chhabra, S.K., Gupta, A.K., Ganguli, A.: Regulation of calcium homeostasis in acute kidney injury: A prospective observational study. Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine **26**(3), 302 (2022)

[28] Thoral, P.J., Peppink, J.M., Driessen, R.H., Sijbrands, E.J., Kompanje, E.J., Kaplan, L., Bailey, H., Kesecioglu, J., Cecconi, M., Churpek, M.,

et al.: Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the amsterdam university medical centers database (amsterdamumcdb) example. Critical care medicine **49**(6), e563 (2021)

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), `http://arxiv.org/abs/1706.03762`

[30] Wan, F., Guo, G., Zhang, C., Guo, Q., Liu, J.: Outlier detection for monitoring data using stacked autoencoder. IEEE Access **7**, 173827–173837 (2019)

[31] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., Sun, L.: Transformers in time series: A survey (02 2022)

[32] Xu, D., Wang, Y., Meng, Y., Zhang, Z.: An improved data anomaly detection method based on isolation forest. In: 2017 10th international symposium on computational intelligence and design (ISCID). vol. 2, pp. 287–291. IEEE (2017)

[33] Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 11121–11128 (2023)

[34] Zimmerman, L.P., Reyfman, P.A., Smith, A.D., Zeng, Z., Kho, A., Sanchez-Pinto, L.N., Luo, Y.: Early prediction of acute kidney injury following icu admission using a multivariate panel of physiological measurements. BMC medical informatics and decision making **19**(1), 1–12 (2019)