

# Multivariable Serum Creatinine Forecasting for Acute Kidney Injury Detection Using an Explainable Transformer-based Model

Cyprien Gille<sup>1\*</sup>, Galaad Altares<sup>2</sup>, Benjamin Colette<sup>3</sup>,  
Karim Zouaoui Boudjeltia<sup>3</sup>, Matei Mancas<sup>1</sup> and Virginie Vandenbulcke<sup>1</sup>

**Abstract**— Acute kidney injury (AKI) in Intensive Care Units (ICU) poses a significant threat to patient health, with potential lasting damage and high mortality risks associated with delayed diagnosis. Serum creatinine (SCr), a key indicator of renal function, is a crucial tool for timely prognosis and treatment. However, SCr values are influenced by numerous complex biological variables that are measured at irregular sparse intervals due to the nature of ICU patient monitoring. This makes algorithmic SCr forecasting particularly challenging.

In this paper, we present an embedding-based interfaced Transformer neural network that effectively handles the complexities of clinical multivariate irregular time series data found in ICU. Our approach demonstrates exceptional performance on SCr forecasting against nine competing models, on two extensive datasets, and across four feature selection scenarios, achieving a median absolute error of 0.043 mg/dL in the most common setting as the first ever model to predict raw SCr values.

To adequately support clinical decision-making, we also emphasize prediction explainability through token-level and feature-level contributions. By providing accurate SCr forecasts, our research has the potential to improve patient outcomes and reduce the risk of AKI-related mortality in ICU settings.

**Clinical relevance**— This provides an accurate prediction of serum creatinine levels which is crucial in ICU where AKI can significantly impact patient outcomes.

## I. INTRODUCTION

This interdisciplinary study aims at forecasting renal failures by developing a reliable, highly competent and flexible framework capable of learning and harnessing the complex dynamics of multivariate irregular time series (MITS). This modality can indeed notably be found in clinical settings such as Intensive Care Units (ICU), where it naturally arises from patient care, and where there is much at stake for decision-assisting information processing systems.

Hereafter, we provide the clinical context and motivation behind our research in section I-A. We then highlight the technical challenges posed by the characteristics of the aforementioned modality in section I-B, before outlining the previous works relevant to our subject matter in section II. We describe our method and the evaluation of its performances in detail in sections III and IV respectively. As decision transparency is essential to any clinical model, we provide an

explainability study of our model in section IV-A. Finally, we discuss the limitations and future perspectives of our work in section V and conclude in section VI.

### A. Medical Context

Acute Kidney Injury (AKI), consisting of an abrupt loss of renal function, is a major cause of adverse health outcomes in ICU. Beside personal symptoms including weakness, vomiting or cramps, AKI also entails undesirable general consequences such as longer stays, higher personal and operating costs [1], and increased mortality [2]. Far from being a rare complication, AKI has been reported to occur in half of ICU patients [3], and is only set to become more frequent in the coming years [4].

Measured via blood sample, Serum Creatinine (SCr) is an essential indicator of kidney health. Indeed, SCr is a byproduct of muscle metabolism which is supposed to be cleared from the blood by the kidneys under regular operation through glomerular filtration. In case of renal failure, SCr blood concentrations will increase. In fact, an increase in SCr levels is one of the criteria (along with reduced urine output) for diagnosing AKI in the widely used KDIGO classification (Kidney Disease: Improving Global Outcomes) [5]. Additionally, SCr concentrations can be used to compute the estimated Glomerular Filtration Rate (eGFR), a widespread tool to evaluate the severity of kidney disease [6].

Concurrently, due to the large variety of sensors, laboratory tests, procedures, fluid inputs and outputs that constitute a typical ICU stay, a considerable amount of clinical information is generated around each intensive care patient. To support clinicians in processing this high-dimensional wealth of data, we intend to leverage modern information processing systems. Ultimately, it would be highly beneficial to intensivists to be provided with a capable, reliable, explainable and robust SCr prediction model, taking advantage of the many measured variables in the ICU setting to anticipate AKI.

### B. Technical Challenges

Multivariate Irregular Time Series (MITS) are not uncommon in real-life settings, as they can appear in domains ranging from geology [7] to behavioral science [8] to environmental forecasting [9]. In the frequent instance of wireless sensor arrays, irregular sampling rates occur due to network fluctuations [10]. Finally and relevantly to our application, any human-operated sensor or data source that

This work was supported by the Walloon region.

<sup>1</sup> C.Gille, M.Mancas and V.Vandenbulcke are with the Information, Signal and Artificial Intelligence (ISIA) Laboratory, University of Mons (UMONS), Mons, Belgium

<sup>2</sup> G.Altares is with the Multitel Research Centre, Mons, Belgium

<sup>3</sup> B.Colette and K.Zouaoui Boudjeltia are with the Experimental Medicine Laboratory (LME), Université Libre de Bruxelles (ULB), Charleroi, Belgium

\* Corresponding author: cyprien.gille@umons.ac.be

acquires information only when prompted naturally produces MITS.

MITS are a significantly more challenging modality to process than classical time series, for a panel of reasons. First, MITS do not have a well-defined sampling frequency. This deprives them of most of the usual time series processing techniques, including all spectral methods, and obfuscates the periodicity of the underlying processes. Second, globally defining the sampling times across all features results in variables with prohibitively high missing rates. Third, distinct features can have greatly different characteristics: varying frequencies of occurrence, ranges, and dynamics. While this is a common hurdle in multivariate problems, it is exacerbated further in the case of MITS, as discrepancy arises not only within the value dimension but also across the temporal dimension. Fourth, feature-level importance is often uncorrelated to the number of occurrences of said feature in the series: in a clinical setting for example, some variables will be measured semi-routinely by an bedside monitoring system, while others will only be manually recorded by physicians. Neither scenario necessarily implies a higher relevance of the feature. Fifth, token-level importance also varies within a feature. This is also a typical aspect of multivariate analysis that is complicated by the sparsity of MITS, as the critical token for a feature might have been observed a long time ago, with numerous observations for other features since then.

Additionally, medical time series have their own specificities, due to the nature of patient care and monitoring [11]. ICU data sources are inherently and materially heterogeneous. The presence and, more challengingly, the missingness of a given feature at a given time has been proven to be clinically informative [12], [13].

## II. RELATED WORK

### A. Creatinine Prediction

Most of the previous research around kidney health forecasting focuses on AKI prediction as a binary classification problem [14], [15], merely predicting whether AKI will occur or not during a set prediction horizon. This approach discards a lot of the nuance surrounding renal function and does not allow the intensivist to interpret the predictions beyond the binary result. Additionally, the rate of false positives often remains too high to meet the quality standard of actual clinical use [16]. Another research area linked to our own work is baseline SCr estimation [17]: the base level of SCr is helpful as a reference point for AKI diagnosis as regular SCr concentrations vary from individual to individual, and are sometimes missing from a patient’s ICU file. Overall, most of the efforts made come from medical research [18]: interdisciplinary research with machine learning expertise has the potential to be fruitful.

Regarding predictor architectures, recent studies include SCr in their inputs and use random forests, gradient boosting, or in rare cases simple recurrent neural networks to perform predictions [15], [19]. A majority also does not take into account the evolution of the patient and are based on summary data, without a temporal dimension [20].

In any case, none of the aforementioned works take up the challenge of forecasting in-ICU raw SCr levels directly: to the best of our knowledge, this study is the first to achieve this.

### B. Multivariate Irregular Time Series

MITS have been the subject of significantly less research than regular time series, but efforts have been made to overcome the challenges laid out in section I-B. Importantly, several strategies have been developed to deal with sampling irregularity. A straightforward approach is discretization [21], where the observed values are aggregated in evenly-spaced temporal bins to obtain a regular time series, which can then be processed using classical methods. With this technique arises the need for imputation strategies to deal with intervals where no value was observed: these can be as simple as forward-filling or zero imputation [22], albeit more complex methods have been developed using generative models [23]. MITS have also been augmented using interpolation, for example through Gaussian Processes [24] or neural networks [25]. Nevertheless, it is important to note that missing data imputation induces biases [26], which is highly undesirable in clinical settings.

More specifically, MITS were successfully processed by several recent studies. First, recurrent networks like Gated Recurrent Units (GRU) [27] were used with hourly aggregation, binary missingness indicators, and time elapsed since the last actual observation as their inputs. They were then improved in by adding a decay cell (GRU-D) [28] that progressively decays the hidden state towards an empirical mean for each time step without an observation. Simultaneously, Ordinary Differential Equations coupled with RNNs (ODE-RNN) [29] allowed for the modeling of continuous-time hidden dynamics. Ideas such as attention were also borrowed from other sequence processing domains, and used to build a multi-timescale model in [30].

Finally, several time representations were developed to circumvent the irregularity issue and multivariate aspect of MITS. Striving for a bias-free representation, [31], [32], [33] all use  $\{value, time, variable\}$  triplets as inputs. To process the triplets, [31] uses set functions, [33] uses attention mechanisms, and [32] uses a graph neural network to model the relationships between the different variables.

## III. METHOD

### A. Model

Hereafter, we present a Transformer-based architecture, interfaced to be able to learn from MITS: the Transformer for Multivariate Irregular Time Series, or T-MITS. The full architecture, which can be used for classification, regression, or forecasting is illustrated by Figure 1.

The input to T-MITS is a sequence of length  $L$  of  $(value, time, variable)$  triplets, which we denote  $\{(x_i, t_i, v_i)\}$  with  $i \in [1 \dots L]$ . This natural triplet representation of MITS was first exploited by [31] and allows for a raw input without any extraneous generated information (unlike missing values imputation methods). Instead of a

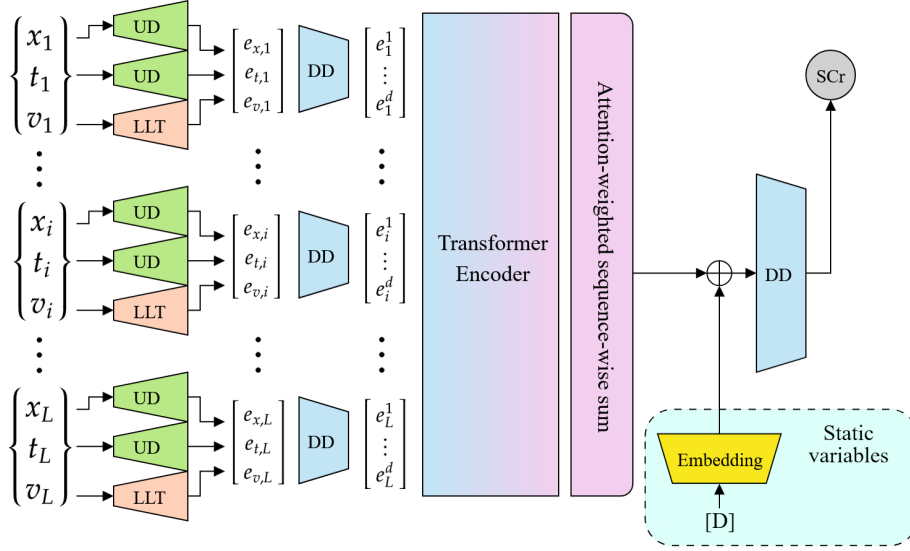


Fig. 1. T-MITS Architecture.  $(x, t, v)$  respectively stand for the observed value, the observation time, and the observed variable. UD indicates an Up-Dimensional ( $d_{in} < d_{out}$ ) feed-forward network (FFN), while DD indicates a Down-Dimensional ( $d_{in} > d_{out}$ ) FFN. LLT indicates a Learnable Lookup Table.  $L$  is the sequence length,  $d$  is the embedding dimension. In our case,  $[D]$  is the demographic information vector.

multivariable misaligned time series, the input becomes a one-dimensional sequence of three-dimensional values, also removing the misalignment hurdles provided that the rest of the architecture is able to process the triplets.

The value and time components of these triplets are processed into embeddings by up-dimensional feed-forward networks (FFN) with hyperbolic tangent activation functions that follow (1). This way of producing continuous input embeddings with one-to-many networks was proven to be effective by [33].

$$\begin{aligned} e_{x,i} &= \text{Linear}_{x,2}(\tanh(\text{Linear}_{x,1}(x_i))) \\ e_{t,i} &= \text{Linear}_{t,2}(\tanh(\text{Linear}_{t,1}(t_i))) \end{aligned} \quad (1)$$

The variable component is embedded using a learnable lookup table, akin to how words are turned into embeddings in natural language processing. All embeddings at this stage are of size  $d$ , an hyperparameter corresponding to the embedding dimension of the model. To combine the value time and variable embeddings, and allow information to flow between the three segments within each token, a down-dimensional FFN processes the concatenation of the three vectors into a single  $d$ -dimensional vector.

The sequence of embeddings is then passed in parallel through several layers of a Transformer encoder [34]. Note that the number of observations varies between stays, which makes the Transformer perfectly suited for our purposes since it can process sequences of arbitrary length. Our implementation differs from the vanilla Transformer in two ways: the sequence order is not included through a sinusoidal positional encoding, but through the direct integration of the time token in each embedded token. Additionally, we swap the ReLU (Rectified Linear Unit) non-linearity for the GELU (Gaussian Error Linear Unit) activation function [35].

Finally, to reduce the sequence along the token dimension, a weighted sum of all tokens is performed with coefficients computed through an FFN also following (1), using the Transformer’s output as its input [33].

Static variables (in our case, patient sex and age) can be integrated at this stage through an embedding model (in our case, another up-dimensional FFN), and then concatenated to the MITS embedding before the final layers of the model. Finally, the resulting vector is processed by one last down-dimensional FFN to obtain the desired output dimension  $O = 1$  for SCr regression. Note that the output dimension can easily be adjusted for classification ( $O = n_{classes}$ ) or forecasting ( $O = n_{variables}$ ).

The architectural choices made above allow T-MITS to process MITS of variable lengths, with an arbitrary number of different variables, with variables entirely missing from a sequence, with long and short inter-variable and intra-variable dependencies.

T-MITS was implemented in Pytorch<sup>1</sup> and is open-sourced under the MIT License, along with all of its training code, trained weights, and preprocessing scripts<sup>2</sup>.

## B. Datasets

For our experiments, we use either the single-center Medical Information Mart for Medical Care IV (MIMIC-IV) [36], [37] or the multi-center eICU Collaborative Research Database (eICU-CRD) [38], which are both publicly available free of charge<sup>34</sup>. Training is required to access the datasets. We use version 2.2 of MIMIC-IV and version 2.0 of eICU-CRD, which are both licensed under the PhysioNet [39]

<sup>1</sup><https://pytorch.org/>

<sup>2</sup><https://github.com/CyprienGille/T-MITS>

<sup>3</sup><https://physionet.org/content/mimiciv/2.2/>

<sup>4</sup><https://physionet.org/content/eicu-crd/2.0/>

TABLE I

EVALUATION METRICS OF ALL MODELS FOR EACH EXPERIMENTAL SETTING. MEANS OVER 5 CROSS-VALIDATION FOLDS, WITH  $2\sigma$  CONFIDENCE INTERVALS. MODEL ABBREVIATIONS ARE DEFINED IN SECTION III-E, AND EXPERIMENTAL SETTINGS ARE DEFINED IN SECTIONS III-B AND IV.

Model	MIMIC 29		eICU		MIMIC 206		MIMIC 206-SCr	
	MedAE $\times$ 100	F1 $\times$ 100	MedAE $\times$ 100	F1 $\times$ 100	MedAE $\times$ 100	F1 $\times$ 100	MedAE $\times$ 100	F1 $\times$ 100
SVR	13.1 $\pm$ 0.4	89.5 $\pm$ 0.3	12.4 $\pm$ 0.4	86.4 $\pm$ 0.8	13.9 $\pm$ 0.2	88.3 $\pm$ 0.6	19.5 $\pm$ 0.6	81.6 $\pm$ 0.4
SGDR	17.8 $\pm$ 0.8	88.0 $\pm$ 0.6	13.8 $\pm$ 1.6	86.1 $\pm$ 1.0	15.0 $\pm$ 0.6	88.8 $\pm$ 0.3	24.8 $\pm$ 1.0	80.3 $\pm$ 0.6
KNR	20.5 $\pm$ 0.4	82.6 $\pm$ 0.6	18.6 $\pm$ 0.8	82.9 $\pm$ 0.6	21.8 $\pm$ 0.8	78.3 $\pm$ 0.6	23.4 $\pm$ 0.8	75.5 $\pm$ 0.4
DTR	11.2 $\pm$ 1.6	89.4 $\pm$ 0.6	12.6 $\pm$ 1.0	85.2 $\pm$ 0.8	11.2 $\pm$ 0.2	88.6 $\pm$ 0.8	22.2 $\pm$ 0.4	77.9 $\pm$ 0.6
GBR	10.1 $\pm$ 0.3	90.8 $\pm$ 0.6	11.4 $\pm$ 0.4	86.7 $\pm$ 0.6	9.8 $\pm$ 0.2	90.3 $\pm$ 0.4	19.1 $\pm$ 0.6	81.7 $\pm$ 0.2
MLPR	15.1 $\pm$ 0.2	88.5 $\pm$ 0.4	13.5 $\pm$ 0.8	85.8 $\pm$ 0.2	18.3 $\pm$ 0.8	85.9 $\pm$ 0.6	28.0 $\pm$ 1.0	78.2 $\pm$ 0.6
GRU-D	8.9 $\pm$ 0.4	91.3 $\pm$ 0.8	10.8 $\pm$ 0.6	85.6 $\pm$ 0.2	8.1 $\pm$ 0.5	91.9 $\pm$ 0.4	18.0 $\pm$ 1.0	82.3 $\pm$ 0.6
TCN	7.5 $\pm$ 0.1	91.6 $\pm$ 0.4	9.4 $\pm$ 1.1	88.9 $\pm$ 0.6	7.2 $\pm$ 0.4	93.1 $\pm$ 0.2	17.8 $\pm$ 0.1	81.9 $\pm$ 0.2
Latent-ODE	6.3 $\pm$ 0.2	93.2 $\pm$ 0.1	8.7 $\pm$ 0.5	88.4 $\pm$ 0.5	5.5 $\pm$ 0.4	93.8 $\pm$ 0.2	17.7 $\pm$ 0.5	82.2 $\pm$ 0.2
T-MITS	<b>4.3 <math>\pm</math> 0.6</b>	<b>94.5 <math>\pm</math> 0.4</b>	<b>8.1 <math>\pm</math> 0.4</b>	<b>90.6 <math>\pm</math> 0.2</b>	<b>4.3 <math>\pm</math> 0.6</b>	<b>94.6 <math>\pm</math> 0.4</b>	<b>16.9 <math>\pm</math> 0.4</b>	<b>83.1 <math>\pm</math> 0.4</b>

Credentialed Health Data License 1.5.0. MIMIC-IV ICU data was collected at the BIDMC hospital in Boston (USA) between 2008 and 2019 through the MetaVision clinical information system, while eICU-CRD data was collected from 208 USA hospitals between 2014 and 2015.

To test both the medical soundness and the discriminative power of our model, we develop two input features selection methods for MIMIC-IV, which we call bottom-up and top-down. Bottom-up feature selection was informed by the medical expertise of intensivists and a comprehensive review of AKI-focused research, and led to a hand-picked subset of 29 vital signs. Top-down feature selection consisted of retaining only the features most present in MIMIC-IV, down to a limit of 50 000 measures across all ICU stays, which leads to 206 selected variables. For eICU-CRD, we use the subset of top-down MIMIC-IV features present in eICU, which leads to 42 selected variables. For brevity, we provide the full lists of features in the T-MITS github repository.

All values and times are normalized with zero mean and unit variance to circumvent range issues. Values are normalized per-variable, and times are normalized globally across all stays of a given set. To prevent data leakage, this step is performed after splitting.

### C. Task

As mentioned above, we tackle a SCr regression task: given an ICU stay up to a certain cutoff time, we train our model to predict the next observed SCr value.

To provide both an easier point of comparison with other works and a more intuitive sense of the performance of the model, we also provide classification metrics through an artificial classification task, obtained with thresholds following the KDIGO criterion [40], [5] as presented in (2). The used thresholds roughly correspond to Normal, Risk, Injury, and Failure kidney states. Note that models trained on the regression task performed better on the artificial classification task than models directly trained as classifiers.

$$\begin{aligned}
 y &= 0 \text{ if } SCr \in [0, 1.35] \\
 y &= 1 \text{ if } SCr \in ]1.35, 2.68] \\
 y &= 2 \text{ if } SCr \in ]2.68, 4.16] \\
 y &= 3 \text{ if } SCr > 4.16
 \end{aligned} \tag{2}$$

### D. Training

We train T-MITS using the Adam optimizer [41] and the Huber loss function [42] described by (3) and chosen for its robustness to outliers. We decay the learning rate by a factor of 0.5 on any loss plateau lasting more than 2 training epochs.

$$L_{\delta}(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{for } |y - y'| \leq \delta, \\ \delta(|y - y'| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{3}$$

All results are obtained through 5-fold cross-validation. Hyperparameters were optimized with Population-Based Bandits [43], implemented in the Ray Tune package [44].

### E. Baselines

We compare T-MITS to all commonly used predictors in medical settings and studies, as well as some other classical models: a Support Vector Regressor (SVR) [45], a Stochastic Gradient Descent Regressor (SGDR) [46], a K-nearest Neighbors Regressor (KNR) [47], a Decision Tree Regressor (DTR) [48], a Gradient Boosting Regressor (GBR) [49], a Multi-Layer Perceptron Regressor (MLPR) [50]. All of these models are implemented in the scikit-learn package [51].

Additionally, we also provide comparisons with more modern architectures, some of which were designed specifically for MITS: a Temporal Convolutional Network (TCN) [52], a Gated Recurrent Unit with Decay (GRU-D) model [28] and a Latent Ordinary Differential Equation (Latent-ODE) model [29]. We refer the reader to the original publications for details on each of these architectures, which were implemented in PyTorch.

Top-9 most relevant variables (Target: 1.70 mg/dL, Prediction: 1.69mg/dL)

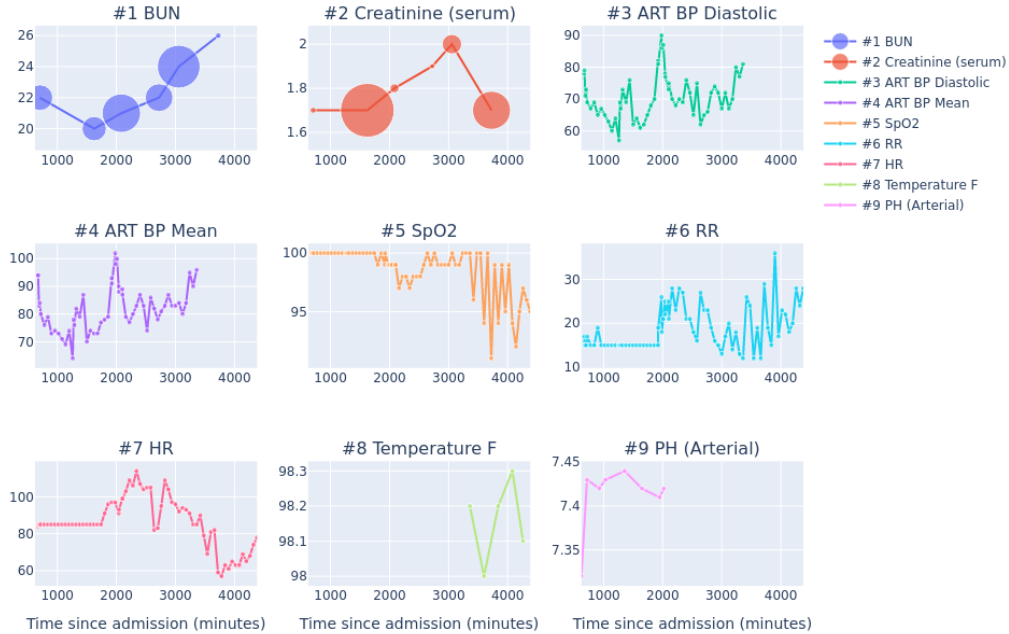


Fig. 2. Visual representation of the top-9 contributing variables in the 29 variables model for an example stay. The size of the token marker is proportional to the absolute value of its attribution. In this case, the BUN and Creatinine variables are the most relevant for the model prediction, while other variables contribute minimally.

#### IV. RESULTS

We present our results in table I. We report the regression Median Absolute Error (MedAE) and the artificial classification F1 score. Note that the MedAE is denormalized, and is thus in mg/dL (for typical SCr values, see (2)).

To paint a more comprehensive picture, we developed two experimental settings on top of the two feature selection methods described in section III-B (denoted in later tables as "29" and "206"). Since SCr is inherently correlated with later values of itself, we follow medical research practices and remove it from the inputs in settings "29-SCr", "206-SCr" and "eICU-SCr". These new settings have two purposes: they provide a more challenging scenario to test our method against, and reveal more nuance in our explainability studies reported in section IV-A. Since the results in these new scenarios are extremely similar between datasets and feature selection methods, we only report them for the "206-SCr" scenario for brevity, but all results are available in our repository with additional applicable metrics (Mean Absolute Error,  $r^2$ , accuracy).

T-MITS consistently outperforms all baselines on all experiments, and reaches very high efficacy on the 29 and 206 settings. Note that for readability, all scores in the results table have been multiplied by 100: T-MITS thus has a median absolute error of 0.043 mg/dL for the set of twenty-nine hand-picked features, with SCr included. As expected, the gap between the different methods shrinks when the provided information is reduced.

##### A. Explainability

To gain more insight into the importance of each variable, we used explainable artificial intelligence (xAI) techniques to extract qualitative representations of attributions for specific samples, and quantitative metrics describing the relative importance of each value in the model prediction. However, MITS as a modality make xAI efforts particularly challenging, as perturbation methods are not suited to the study thereof. Moreover, it is of note that some classes of xAI methods are less effective on a Transformer Encoder, as reported by [53] and [54].

For these reasons, we chose the Shapley Value Sampling method introduced by [55] that is readily applicable in this use case. Indeed, it relies on computing the marginal contribution of tokens by removing them from the input, which is straightforward in our case as we can simply remove a token from the sequence without introducing artifacts (as one would with a black pixel in the image case, or by interpolating missing values in evenly sampled time series). Moreover, Shapley Value Sampling is a landmark xAI method for regression, and is used as a baseline for computing the quality of other xAI methods in regression tasks [56].

In Figure 2, we show an example of the token-level results given by xAI, where it appears that the attributions of input tokens are largely concentrated among 2 variables for the "29" experiment: Blood Urea Nitrogen (BUN) and Serum Creatinine values (while the rest of the variables are general health indicators such as Arterial Blood Pressure or Heart

Rate). This result is consistent with the biological workings of renal function, as urea is also filtered by the glomerulus [57].

Aggregating the contributions of each variable across all tokens and all samples confirms these attributions: the model primarily uses SCr and BUN. However, when given more variables than the 29 hand-picked ones, it also takes advantage of other kidney-related variables, such as the presence of a dialysis catheter, to improve its performance. This is especially revealed when SCr is removed from the input.

## V. DISCUSSION

The main limitation of our work is the fact that SCr remains a late marker of AKI - even with its accurate forecast over a long prediction horizon, T-MITS could provide even better and more useful results using metabolites for example [58]. Another future direction for our work will be to demonstrate the transferability of our learned model to a local in-house database. Finally, we are also working on a trial to actually deploy our Transformer for Multivariate Irregular Time Series (T-MITS) in the field and integrate it in the intensivist's set of decision-assisting tools.

## VI. CONCLUSION

The T-MITS architecture demonstrates robust capability in processing irregular, sparse, misaligned, multivariate time series, and as such is perfectly suited to efficiently leverage the abundance of data produced in intensive care settings. The ability to predict SCr levels can greatly help in treating AKI preventively, a critical effort in the ICU considering its various negative impacts on patient health and hospital load.

Additionally, we showed that the T-MITS architecture could lend itself well to XAI efforts, usually complicated by the specificities of MITS-processing models. This is indispensable for clinical usability. Our versatile and open-sourced modular framework can also easily be applied to a multitude of other natural use cases involving MITS.

## REFERENCES

- [1] S. A. Silver and G. M. Chertow, "The Economic Consequences of Acute Kidney Injury," *Nephron*, vol. 137, pp. 297–301, June 2017.
- [2] E. J. See, K. Jayasinghe, N. Glassford, M. Bailey, D. W. Johnson, K. R. Polkinghorne, N. D. Toussaint, and R. Bellomo, "Long-term risk of adverse outcomes after acute kidney injury: A systematic review and meta-analysis of cohort studies using consensus definitions of exposure," *Kidney International*, vol. 95, pp. 160–172, Jan. 2019.
- [3] C. Ronco, R. Bellomo, and J. A. Kellum, "Acute kidney injury," *The Lancet*, vol. 394, pp. 1949–1964, Nov. 2019.
- [4] J. R. Brown, M. E. Rezaee, E. J. Marshall, and M. E. Matheny, "Hospital Mortality in the United States following Acute Kidney Injury," *BioMed Research International*, vol. 2016, no. 1, p. 4278579, 2016.
- [5] Kidney Disease: Improving Global Outcomes (KDIGO) Glomerular Diseases Work Group, "KDIGO 2021 Clinical Practice Guideline for the Management of Glomerular Diseases," *Kidney International*, vol. 100, pp. S1–S276, Oct. 2021.
- [6] A. M. Cusumano, C. Tzanno-Martins, and G. J. Rosa-Diez, "The Glomerular Filtration Rate: From the Diagnosis of Kidney Function to a Public Health Tool," *Frontiers in Medicine*, vol. 8, p. 769335, Nov. 2021.
- [7] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths, "Comparison of correlation analysis techniques for irregularly sampled time series," *Nonlinear Processes in Geophysics*, vol. 18, pp. 389–404, June 2011.
- [8] T. Ruf, "The Lomb-Scargle Periodogram in Biological Rhythm Research: Analysis of Incomplete and Unequally Spaced Time-Series," *Biological Rhythm Research*, Apr. 1999.
- [9] A. S. Nejad, R. Alaiz-Rodríguez, G. D. McCarthy, B. Kelleher, A. Grey, and A. Parnell, "SERT: A transformer based model for multivariate temporal sensor data with missing values for environmental monitoring," *Computers & Geosciences*, vol. 188, p. 105601, June 2024.
- [10] Y.-F. Zhang, P. J. Thorburn, W. Xiang, and P. Fitch, "SSIM—A Deep Learning Approach for Recovering Missing Time Series Sensor Data," *IEEE Internet of Things Journal*, vol. 6, pp. 6618–6628, Aug. 2019.
- [11] S. N. Shukla and B. M. Marlin, "A Survey on Principles, Models and Methods for Learning from Irregularly Sampled Time Series," Jan. 2021.
- [12] R. H. H. Groenwold, "Informative missingness in electronic health record systems: The curse of knowing," *Diagnostic and Prognostic Research*, vol. 4, p. 8, Dec. 2020.
- [13] A. Sharafoddini, J. A. Dubin, D. M. Maslove, and J. Lee, "A New Insight Into Missing Data in Intensive Care Unit Patient Profiles: Observational Study," *JMIR medical informatics*, vol. 7, p. e11605, Jan. 2019.
- [14] Q. Qian, J. Wu, J. Wang, H. Sun, and L. Yang, "Prediction Models for AKI in ICU: A Comparative Study," *International Journal of General Medicine*, vol. 14, pp. 623–632, 2021.
- [15] T. Ozrazgat-Baslanti, T. J. Loftus, Y. Ren, M. M. Ruppert, and A. Bihorac, "Advances in artificial intelligence and deep learning systems in ICU related AKI," *Current opinion in critical care*, vol. 27, pp. 560–572, Dec. 2021.
- [16] A. Kamel Rahimi, M. Ghadimi, A. H. van der Vegt, O. J. Canfell, J. D. Pole, C. Sullivan, and S. Shrapnel, "Machine learning clinical prediction models for acute kidney injury: The impact of baseline creatinine on prediction efficacy," *BMC medical informatics and decision making*, vol. 23, p. 207, Oct. 2023.
- [17] E. Ghosh, L. Eshelman, S. Lanius, E. Schwager, K. S. Pasupathy, E. F. Barreto, and K. Kashani, "Estimation of Baseline Serum Creatinine with Machine Learning," *American Journal of Nephrology*, vol. 52, no. 9, pp. 753–762, 2021.
- [18] F. Alfieri, A. Ancona, G. Tripepi, D. Crosetto, V. Randazzo, A. Paviglianiti, E. Pasero, L. Vecchi, V. Cauda, and R. M. Fagugli, "A deep-learning model to continuously predict severe acute kidney injury based on urine output changes in critically ill patients," *Journal of Nephrology*, vol. 34, pp. 1875–1886, Dec. 2021.
- [19] Y. H. Du, C. J. Guan, L. Y. Li, and P. Gan, "Predictive value of machine learning for the risk of acute kidney injury (AKI) in hospital intensive care units (ICU) patients: A systematic review and meta-analysis," *PeerJ*, vol. 11, p. e16405, Nov. 2023.
- [20] T. H. Lee, J.-J. Chen, C.-T. Cheng, and C.-H. Chang, "Does Artificial Intelligence Make Clinical Decision Better? A Review of Artificial Intelligence and Machine Learning in Acute Kidney Injury Prediction," *Healthcare*, vol. 9, p. 1662, Dec. 2021.
- [21] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, Apr. 2019.
- [22] Z. C. Lipton, D. Kale, and R. Wetzel, "Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series," in *Proceedings of the 1st Machine Learning for Healthcare Conference*, pp. 253–270, PMLR, Dec. 2016.
- [23] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing Data Imputation using Generative Adversarial Nets," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 5689–5698, PMLR, July 2018.
- [24] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, Nov. 2005.
- [25] S. N. Shukla and B. M. Marlin, "Interpolation-Prediction Networks for Irregularly Sampled Time Series," Sept. 2019.
- [26] B. Y. Gravesteijn, E. W. Steyerberg, and H. F. Lingsma, "Modern Learning from Big Data in Critical Care: Primum Non Nocere," *Neurocritical Care*, vol. 37, no. Suppl 2, pp. 174–184, 2022.
- [27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," Dec. 2014.
- [28] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," *Scientific Reports*, vol. 8, p. 6085, Apr. 2018.
- [29] Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud, "Latent Ordinary Differential Equations for Irregularly-Sampled Time Series," in *Ad-*

- vances in *Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.
- [30] S. N. Shukla and B. M. Marlin, “Multi-Time Attention Networks for Irregularly Sampled Time Series,” June 2021.
- [31] M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt, “Set functions for time series,” in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *ICML’20*, pp. 4353–4363, JMLR.org, July 2020.
- [32] X. Zhang, M. Zeman, T. Tsiligkaridis, and M. Zitnik, “Graph-Guided Network for Irregularly Sampled Multivariate Time Series,” in *International Conference on Learning Representations*, Oct. 2021.
- [33] S. Tipirneni and C. K. Reddy, “Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series,” *ACM Transactions on Knowledge Discovery from Data*, vol. 16, pp. 105:1–105:17, July 2022.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, (Red Hook, NY, USA), pp. 6000–6010, Curran Associates Inc., Dec. 2017.
- [35] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” June 2023.
- [36] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, “MIMIC-IV (version 2.2),” 2023.
- [37] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, “MIMIC-IV, a freely accessible electronic health record dataset,” *Scientific Data*, vol. 10, p. 1, Jan. 2023.
- [38] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, “The eICU Collaborative Research Database, a freely available multi-center database for critical care research,” *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.
- [39] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, pp. E215–220, June 2000.
- [40] S. Calvert and A. Shaw, “Perioperative acute kidney injury,” *Perioperative Medicine*, vol. 1, p. 6, Dec. 2012.
- [41] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *CoRR*, Dec. 2014.
- [42] P. J. Huber, “Robust Estimation of a Location Parameter,” *The Annals of Mathematical Statistics*, vol. 35, pp. 73–101, Mar. 1964.
- [43] J. Parker-Holder, V. Nguyen, and S. J. Roberts, “Provably Efficient Online Hyperparameter Optimization with Population-Based Bandits,” in *Advances in Neural Information Processing Systems*, vol. 33, pp. 17200–17211, Curran Associates, Inc., 2020.
- [44] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, “Tune: A Research Platform for Distributed Model Selection and Training,” July 2018.
- [45] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, May 2011.
- [46] L. Bottou, “Online algorithms and stochastic approximations,” in *Online Learning and Neural Networks*, Cambridge, UK: Cambridge University Press, 1998.
- [47] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, Jan. 1967.
- [48] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Taylor & Francis, Jan. 1984.
- [49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [50] G. E. Hinton, “Connectionist learning procedures,” *Artificial Intelligence*, vol. 40, pp. 185–234, Sept. 1989.
- [51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [52] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal Convolutional Networks for Action Segmentation and Detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012, IEEE Computer Society, July 2017.
- [53] S. Abnar and W. Zuidema, “Quantifying Attention Flow in Transformers,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 4190–4197, Association for Computational Linguistics, July 2020.
- [54] H. Chefer, S. Gur, and L. Wolf, “Transformer Interpretability Beyond Attention Visualization,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, June 2021.
- [55] E. Strumbelj and I. Kononenko, “An Efficient Explanation of Individual Classifications using Game Theory,” *The Journal of Machine Learning Research*, vol. 11, pp. 1–18, Mar. 2010.
- [56] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, “Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective,” *IEEE Signal Processing Magazine*, vol. 39, pp. 40–58, July 2022.
- [57] N. Baum, C. C. Dichoso, and C. E. Carlton, “Blood urea nitrogen and serum creatinine. Physiology and interpretations,” *Urology*, vol. 5, pp. 583–588, May 1975.
- [58] C. H. Johnson, J. Ivanisevic, and G. Siuzdak, “Metabolomics: Beyond biomarkers and towards mechanisms,” *Nature reviews. Molecular cell biology*, vol. 17, pp. 451–459, July 2016.