Predicting Acute Kidney Injury via Interpretable Ensemble Learning and Attention Weighted Convoutional-Recurrent Neural Networks

Yu-Chung Peng*, Niharika Shimona D'Souza†, Brian Bush‡, Charles Brown‡, Archana Venkataraman †
*Department of Computer Science, Johns Hopkins University, Baltimore, MD
†Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD
‡ Department of Anesthesiology and Critical Care Medicine, Johns Hopkins School of Medicine, Baltimore, MD

2021 55th Annual Conference on Information Sciences and Systems (CISS) | 978-1-6654-1268-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISS50987.2021.9400242

***Abstract*—Acute Kidney Injury (AKI) is one of the most frequent postoperative complications and is associated with both short- and long-term mortality. Improved prediction of AKI is crucial and may help clinicians prevent and mitigate its adverse effects. In this paper, we explore the use of machine learning methods to predict postoperative AKI. Our analysis centers on the ensemble-based random forest (RF) classifier, which operates on static clinical variables, and a novel deep learning architecture that incorporates intraoperative time series data along with the static variables. The architecture uses a dual-attention mechanism to select both features and time intervals relevant for AKI prediction. We evaluate our models on the publicly available VitalDB database of 3,640 patients who underwent non-cardiac surgery. The RF outperformed existing machine learning classifiers in the AKI literature (AUROC: 0.86, AUPRC: 0.54). In addition, the RF identified a robust set of preoperative variables that can be screened in a simple blood test. While the deep learning model achieved slightly lower performance (AUROC: 0.84, AUPRC: 0.44), the attention weights provide important intraoperative information, which can be monitored by clinicians during surgery. Taken together, our results highlight the promise of machine learning for AKI prediction and take the first steps towards developing clinically translatable models.**

***Index Terms*—Acute Kidney Injury, Random Forest, Convolutional Neural Network, Recurrent Neural Network**

I. Introduction

Acute Kidney Injury (AKI) is a common post-surgical complication that refers to an abrupt decline in renal function, characterized by retention of nitrogenous waste products and creatinine, along with dysregulation of extracellular fluids. AKI occurs in approximately 5–7.5% of all acute care hospitalizations, and roughly 30–40% of these cases are observed in conjunction with surgery [1]. AKI is a serious morbidity associated with longer hospital stays, increased mortality, and greater risk of developing long-term chronic kidney injury [2], [3]. Hence, there is a clear need for perioperative physicians to understand the baseline and ongoing risk for AKI. Such information could help guide multiple aspects of perioperative management, including drug dosing, fluid titration, hemodynamic targets, and monitoring strategies [4], [5].

The risk for AKI is difficult to predict in any individual patient, as it can involve multiple patient comorbidities and surgical factors. As a consequence, the current standard of care relies heavily on (subjective) expert assessment, clinical expertise, and simple risk-prediction models based on patient characteristics. The rise of machine learning has fueled predictive models for AKI that leverage classical techniques, such as logistic regression, support vector machines, and random forests [6]. Going one step further, the seminal work by [7] introduced a deep learning architecture based on recurrent neural networks to forecast AKI based on patient history. While these methods provide a valuable starting point, they have three key limitations. First, existing studies focus on well-known preoperative variables (demographics, comorbidities, lab values) but do not leverage intraoperative information. Intraoperative data can help illuminate physiological insults that occur during surgery and are thought to play a key role in the development of AKI. In addition, many of these intraoperative variables can be modified during the surgery, thus providing a crucial opportunity to *prevent* AKI before it develops. Second, deep learning methods in particular provide black-box predictions, which make it difficult to disentangle the relevant influences. This is a major detriment to developing better clinical management strategies for AKI. Third, prior work focuses on just two evaluation metrics: prediction accuracy and area under the receiver operating characteristic curve (AUROC). Both metrics are overly optimistic in the case of severe class imbalance [8], which makes it unlikely that the methods will generalize to larger clinical populations.

In this paper, we take a critical approach to the design, application, and evaluation of machine learning for AKI prediction that addresses the above limitations. First, we propose a novel deep learning architecture to integrate static clinical variables with intraoperative time series data. This architecture uses a convolutional neural network (CNN) to construct a low-dimensional encoding [9] and a long short-term memory (LSTM) module to capture temporal dependencies. The CNN-LSTM is complemented by a static random forest (RF) classifier [10], which achieves one of the highest AKI prediction performances reported to date. Second, both models provide interpretable feature importance scores. For the CNN-LSTM, these scores are computed by a novel dual attention module that identifies both clinical variables and intraoperative time intervals that predict AKI. Third, we propose a larger suite

This work was supported by the NSF CAREER award 1845430 and by the Malone Center for Engineering and Healthcare seed grant.

of evaluation metrics, including the area under the precision-recall curve (AUPRC), which emphasizes the minority AKI class. Finally, we evaluate our models on a large publicly available database of heterogeneous surgical cases. Hence, our approach can be easily replicated and provides an important benchmark for future AKI research.

II. Clinical Dataset

A. Data Source

Data used in this study was drawn from the VitalDB database, which contains preoperative and intraoperative data from 6,388 patients who underwent non-cardiac surgery [11]. The database was compiled and released by the VitalLab in 2016 and consists of patients admitted to the Seoul National University Hospital in Seoul, Republic of Korea. VitalDB contains patient comorbidities, preoperative laboratory values, surgical factors, and intraoperative signals sampled at 1–7 seconds. VitalDB also contains postoperative measurements, which we use to determine the incidence of AKI. For this study, we use a criteria outlined by the KDIGO (Kidney Disease: Improving Global Outcomes) organization: an absolute increase of creatinine levels by 0.3mg/dL or an increase to 1.5 times the patient's baseline levels defines AKI [12]. Due to data limitations, we were unable to identify and exclude patients on dialysis. Only patients with both a baseline creatinine value at least one postoperative creatinine value were included in this study.

B. Data Processing

Data was downloaded through the VitalDB API in the form of csv files. We also subselected 16 intraoperative time series that were acquired for the majority of patients: body temperature, diastolic/systolic blood pressure, end-tidal CO_2, fraction of inspired/expired N_2O, fraction of inspired/expired O_2, heart rate, propofol infusion rate, infused volume of propofol, remifentanil infusion rate, infused volume of remifentanil, respiratory rate, percutaneous O_2 saturation, and tidal volume.

Static variables that were missing in more than 30% of patients were removed, resulting in 62 features for analysis. These features included demographics, like age and sex, preoperative lab test baselines, like sodium and hemoglobin, and intraoperative summaries, like operation position and duration. Categorical variables were encoded using label encoding, and missing values were imputed using the K-nearest neighbors algorithm. Lab test data was used exclusively to diagnose a patient with AKI using the KDIGO criteria to create binary labels for each patient, with 1 representing that the patient developed AKI in the next 7 days (the critical monitoring period) and 0 representing that the patient did not develop AKI within this time. We resampled the intraoperative data to 0.1 Hz and linearly interpolated missing values from the neighboring time points. To address the varying lengths of operations, time points were replicated from shorter operations until all samples were of the same length. Finally, patients with operation durations in the top and bottom 2% were removed. In total, our final cohort contained 3,640 patients. Table I shows demographic and clinical information about the patient cohort.

TABLE I
Statistics of the VitalDB data used in this study.

Category	*AKI*	*No-AKI*	*Total*
Sex(Male)	149	1961	2110
Age(Mean)	52.8	59.9	59.5
BMI(Mean)	22.3	23.0	22.9
General surgery	205	2515	2720
Thoracic	18	770	788
Gynecology	1	52	53
Urology	7	72	79
Total	231	3409	3640

C. Class Imbalance

We note that only 231 out of the 3640 patients (6.35%) had a record of developing postoperative AKI. This massive class imbalance makes it difficult to evaluate model performance. Specifically, a classifier can achieve high accuracy simply by predicting the label 0 (no AKI) for all patients. Thus, we train the models using a weighted loss in which wrongly classifying an AKI patient will incur a greater penalty than the reverse.

III. Methods

A. Random Forest Model

Random forest (RF) is an ensemble learning method which combines bagging decision trees, grown across random subsets of the data, with random feature sub-sampling to construct the decision splits. The dual randomization has been shown to reduce model overfitting since features are chosen individually by each decision tree, and results are aggregated based on multiple such data splits. Statistically speaking, this bagging procedure helps mitigate errors made by individual trees in the ensemble to infer a more robust feature selection. In this vein, the RF computes a selection weight for each feature known as the Gini Importance. At a high level, GI quantifies how well a particular feature separates the two classes, across all associated decision nodes in the forest.

Our RF is composed of 400 trees with a max depth of 27 using class weights inversely proportional to their frequency. The input to the RF was a vector of static variables (demographics, comorbidities, surgical factors), and the task was a binary classification of AKI versus no AKI.

B. Deep Learning Architecture

In parallel to the RF, we aim to integrate the static variables with the intraoperative data space. As the RF does not provide a natural way to incorporate dynamics, we approach this problem from the deep learning lens. Fig. 1 illustrates our framework. At a high level, our model emphasizes prominent information from the intraoperative data using a CNN (green block in Fig. 1) with both channel and temporal attention modules (Figs. 2 and 3) and feeds it to a recurrent neural network for the final prediction (purple block in Fig. 1).

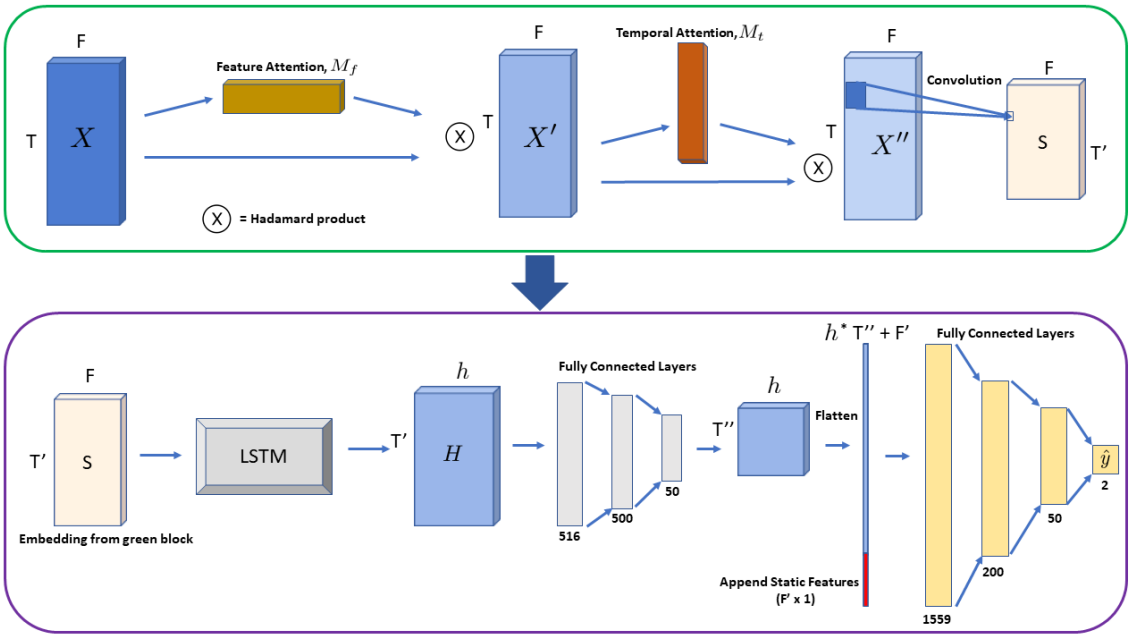

Fig. 1. Overview of our CNN-LSTM architecture for postoperative AKI prediction. **Green block:** Convolutional model for feature aggregation. Given the intraoperative data, it detects pertinent features and re-weights the input data. **Purple block:** Recurrent portion for modeling temporal patterns throughout operations. The input for this block is the obtained convolutional embedding $\mathbf{S}$ from the Green block.

CNN for feature extraction: The CNN portion of our model was inspired by the CBAM developed in [13], which implements two attention maps corresponding to kernel-wise and spatial descriptors. To cater to our application, we modify this setup such that the attention is computed both across time, as well as across features of the intraoperative data.

Mathematically, let $\mathbf{X} \in \mathcal{R}^{F \times T}$ be the intraoperative data for a single patient, where F is the number of features and T is the number of time steps. Our model creates a 1D feature attention vector $\mathbf{M}_f \in \mathcal{R}^{F \times 1}$ and a 1D temporal attention vector $\mathbf{M}_t \in \mathcal{R}^{1 \times T}$. The overall attention weighted feature extraction operation can be formalized as

$$\mathbf{X}'(:, c) = \mathbf{M}_f \otimes \mathbf{X}(:, c) \tag{1}$$

$$\mathbf{X}''(r, :) = \mathbf{M}_t \otimes \mathbf{X}'(r, :) \tag{2}$$

where $\otimes$ indicates the element-wise Hadamard product. At a high level, we wish to underscore features that are most relevant to prediction in $\mathbf{M}_f$. Simultaneously, $\mathbf{M}_t$ is designed to highlight temporal events during the surgery that are important for classification.

Feature Attention Module: Fig. 2 illustrates the construction of our feature-wise attention module. To obtain the feature attention vector, the data is first passed into two pooling layers. The first of these performs a max pooling operation, giving rise to $\mathbf{P}^f_{\text{max}} \in \mathcal{R}^{F \times 1}$. The second computes a mean pool, resulting in $\mathbf{P}^f_{\text{mean}} \in \mathcal{R}^{F \times 1}$. The two vectors are then input to two fully connected layers. From here, they are added together and passed through the sigmoid function to obtain $\mathbf{M}_f$:

$$\mathbf{M}_f = \sigma[\mathbf{\Phi}_1(\mathbf{\Phi}_0(\mathbf{P}^f_{\text{mean}})) + \mathbf{\Phi}_1(\mathbf{\Phi}_0(\mathbf{P}^f_{\text{max}}))], \tag{3}$$

Here, σ represents the sigmoid function, and $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_1$ are fully connected linear layers with ReLU activation. Eq. (3) mimics an encoder-decoder architecture designed for data-compression, as shown in Fig. 2. This parametrizes a learnable transformation to accentuate directions in the native data-space that aid the AKI classification task.

Temporal Attention Module: Our temporal attention network is depicted in Fig. 3. Similar to the feature attention, we use a simple neural network to learn the temporal attention weight vector $\mathbf{M}_t$. Again, the data is input to two pooling layers. The first of these constructs the max pooled output $\mathbf{U}^t_{\text{max}} \in \mathbf{R}^{1 \times T}$, while the second performs a mean pool to give $\mathbf{U}^t_{\text{mean}} \in \mathbf{R}^{1 \times T}$. The two vectors are then concatenated along the row. From here, they feed into a convolutional layer with a sigmoid activation to learns the attention weights. Mathematically,

$$\mathbf{M}_t = \sigma(\mathcal{F}_{\text{conv}}([\mathbf{U}^t_{\text{max}}; \mathbf{U}^t_{\text{max}}])), \tag{4}$$

Here, $\mathcal{F}_{\text{conv}}$ parametrizes a convolution layer with tunable filter weights learned via backpropagation.

LSTM-ANN for Classification: Referring to Fig. 1, $\mathbf{S} \in \mathcal{R}^{F \times T}$ represents the attention-weighted data embedding obtained from the CNN. We feed this representation into a recurrent neural network (RNN) that models the temporal evolution. We chose the Long Short-Term Memory (LSTM) variant, which is known for its improved convergence properties, while faithfully modeling temporal trends [14]. We opted for an LSTM with 2 hidden layers of size h = 30. This hidden representation $\mathbf{H} \in \mathcal{R}^{h \times T'}$ from the LSTM is further input to a three layered fully connnected network with ReLU

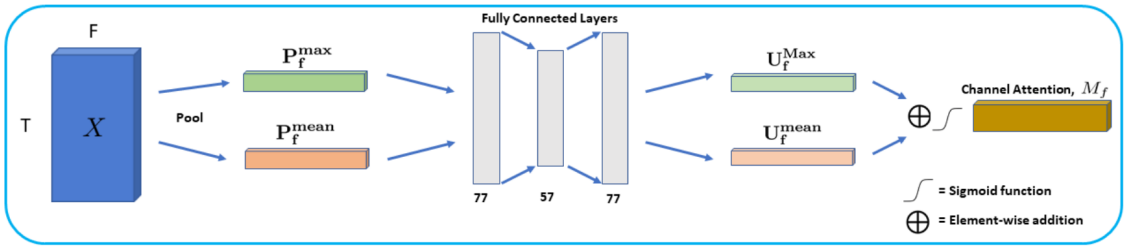

Fig. 2. Feature attention to select and amplify important features.

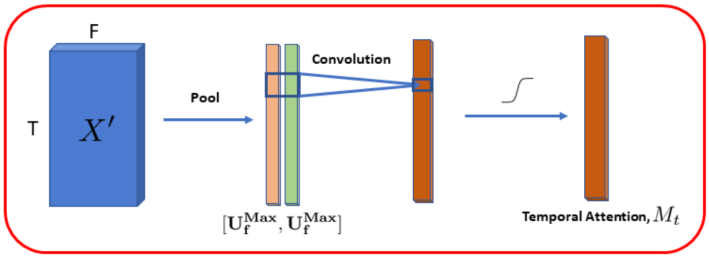

Fig. 3. Temporal attention to select and amplify important time points.

activations and batch normalization and output dimension T" to obtain $\hat{\mathbf{S}} \in \mathcal{R}^{h \times T''}$. At this stage, we introduce the pre-operative data features by appending them to a flattened version of $\hat{\mathbf{S}}$. Finally, the augmented representation is passed to a fully connected classifier comprised of three hidden layers with ReLU activations and batch normalization. The resulting final output is a two dimensional vector $\hat{\mathbf{y}}$ which encodes probabilities that a patient develops AKI vs not.

Implementation Details: The model was trained to minimize a weighted cross entropy loss with weight ratio of $0.05 : 0.95$ for patients characterized as non-AKI vs AKI. The entire deep network was trained end-to-end with the ADAM [15] algorithm using a learning rate of $8 * 10^{-4}$, a weight decay of 0.0001, and a batch size of 32 on a NVIDIA 2070 GPU. The intraoperative dataset was modified such that static clinical variables were appended to each time point to simultaneously provide historical and intraoperative context to the network.

C. Baseline Methods

We compare the RF and CNN-LSTM models to two classical machine learning methods: support vector machine (SVM), and logistic regression (LR). Both methods operate on just the static variables. Once again, the task was binary classification of AKI versus no AKI, denoted as "1" and "0", respectively.

1) SVM: The Support Vector Machine (SVM) classifier constructs hyperplanes in a potentially complex feature space to separate input points into its constituent classes. Formally, these hyperplanes are constructed such that the separation between the two classes is maximized [16]. We opt for a linear kernel, which we found empirically to be more robust than nonlinear variants along with a hinge loss penalty C=1.

2) Logistic Regression: Logistic regression is a popular statistical technique used to model a binary dependent variable using the 'logistic' function [10]. This procedure constructs a line of best fit through the input data points to estimate a probability measure for developing AKI vs not. Finally, for our implementation, we add an ℓ_2 regularization on the regression coefficients with a penalty factor of $C = 1$

D. Evaluation Strategy

We use 5-fold cross validation, run across 10 different data splits, to quantify the performance of each model. The validation metrics are averaged across all testing folds.

Most prior work on AKI prediction report just the area under the receiving operator curve (**AUROC**), which evaluates the true positive rate (TPR) versus the false positive rate (FPR) as the detection threshold is varied. Mathematically,

$$TPR = \frac{TP}{TP + FN} \quad \text{and} \quad FPR = \frac{FP}{FP + TN} \tag{5}$$

where TP (true positive) is the number of times the classifier correctly classifies a patient with a true label of 1, TN (true negative) is the number of times the classifier correctly classifies a 0, FN (false negative) is the number of times the classifier predicts 0 when the true label is 1, and FP (false positive) is when the classifier predicts 1 when the true label is 0. However, AUROC is an overly optimistic metric in cases with extreme class imbalance, such as our VitalDB dataset. For example, an algorithm can generate several false positive predictions in an attempt to "guess" the true positive. Since the majority class is large, this trade off will have minimal impact on the FPR, resulting in a high overall AUROC. Hence, we argue for the area under the precision recall curve (**AUPRC**) to be the standard metric for this application [17]. Precision

TABLE II
QUANTITATIVE PERFORMANCE OF ALL MODELS USING REPEATED 5-FOLD CROSS VALIDATION. HIGHER AUROC AND AUPRC INDICATE BETTER PERFORMANCE; LOWER STANDARD DEVIATION INDICATES ROBUSTNESS.

Model	*AUROC*	*AUPRC*	*F1-Score*	*Sensitivity*	*Specificity*	*Accuracy*
Logistic Regression	0.83 $\pm$.007	0.34 $\pm$.008	0.33 $\pm$.088	0.43 $\pm$.119	0.95 $\pm$.016	0.91 $\pm$.008
SVM	0.83 $\pm$.008	0.33 $\pm$.010	0.31 $\pm$.055	0.40 $\pm$.077	0.95 $\pm$.012	0.91 $\pm$.007
CNN-LSTM	0.84 $\pm$.013	0.44 $\pm$.018	0.39 $\pm$.055	0.38 $\pm$.057	0.97 $\pm$.008	0.94 $\pm$.005
RF	0.87 $\pm$.006	0.54 $\pm$.011	0.55 $\pm$.009	0.50 $\pm$.018	0.98 $\pm$.003	0.95 $\pm$.002

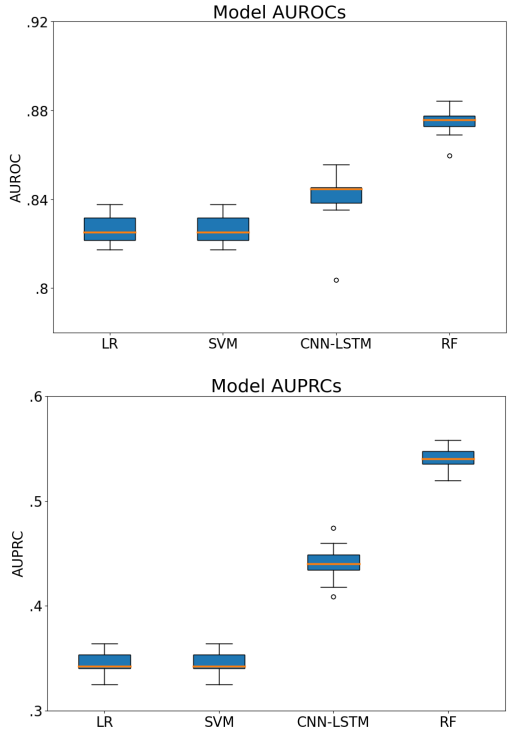

Fig. 4. **Top:** Boxplots of different models' AUROCs evaluated over 10, 5-fold cross validation runs. **Bottom:** Boxplots of different models' AUPRCs evaluated over 10, 5-fold cross validation runs.

captures the ratio of true positive to false positive detections, thus disabmiguating the previous situtaion. Formally,

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{and} \quad \text{Recall} = \frac{TP}{TP+FN} \tag{6}$$

Finally, we report the F1 score, which is the harmonic mean of Precision and Recall. For metrics requiring a threshold (e.g., TPR and FPR), we pick the threshold based on the value that results in the maximum f1-score, discussed in [18].

IV. RESULTS

Table II reports the quantitative performance of all four models across repeated 5-fold cross validation runs. As seen, the RF achieves the highest AUROC and AUPRC values, indicating that the static clinical variables carry most of the information for AKI prediction. We note that the CNN-LSTM achieves also achieves significantly higher AUPRC than logistic regression ($p < 0.0001$) and the SVM ($p < 0.0001$).

TABLE III
RF FEATURE IMPORTANCE SCORES

Feature	*Mean*	*Std* (10^{-3})
Operation type	0.130	6.26
Preoperative albumin	0.049	3.20
Preoperative hemoglobin	0.044	3.54
Preoperative creatinine	0.031	2.17
Age	0.028	2.55
Preoperative blood urea nitrogen	0.027	2.73
Intraoperative estimated blood loss	0.027	2.32
Preoperative platelet count	0.027	1.87

While the CNN-LSTM achieves higher mean AUROC than both baselines, the difference is not significant for SVM. The CNN-LSTM also had a significantly higher F1 score than the SVM ($p < 0.0018$) and logistic regression ($p < 0.041$) and a significantly higher specificity and accuracy compared to both models ($p < 0.0001$ for all comparisons). These statistics highlight the weakness of AUROC as a holistic evaluation metric when the CNN-LSTM, with similar AUROC to the LR and SVM, produces only two thirds of the errors.

Table III lists the top clinical variables selected by the RF, along with the corresponding Gini Importance. It is noteworthy that the majority of the top features are patient risk factors, such as age and preoperative laboratory values. Only one intraoperative value (estimated blood loss) was included in the model. These results suggest that baseline patient risk factors, to a greater degree than intraoperative events, are primary risk factors for AKI. These results support the importance of patient selection and/or preoperative strategies that might improve optimization prior to surgery.

Fig. 5 displays the attention weights learned by the CNN-LSTM for 100 representative patients. The first 50 rows correspond to patients who developed AKI and the last 50 rows correspond to patients who did not develop AKI.

The feature attention weights (top of Fig. 5) are designed to select the relevant clinical variables for each patient. They are organized such that the intraoperative time series features are to the left of the black line, and the (static) preoperative and demographic variables are to the right. Although the feature attention is slightly different for each patient, we note that the most consistently highlighted features are fraction of inspired oxygen, heart rate, and remifentanil rate for the time series variables, and preoperative albumin, case duration, and preoperative creatinine for the static variables. Notice that they are consistently selected in both the AKI and non-

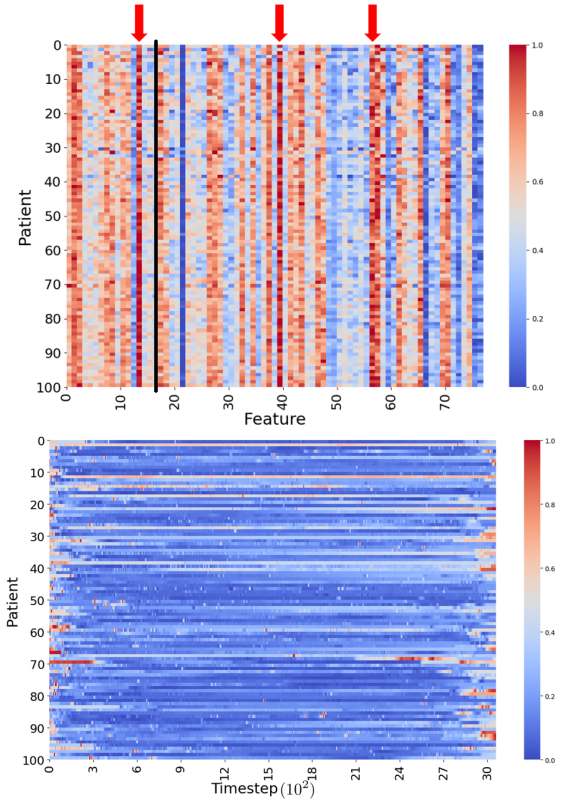

Fig. 5. **Top:** Feature attention weights, where clinical variables are organized along the x-axis and patients along the y-axis. Warmer colors indicate a greater weight that the model puts on the corresponding feature. Red arrows point to intraoperative remifentanil volume, Preoperative diabetes, and preoperative creatinine as examples of commonly selected features. **Bottom:** Temporal attention weights, where time proceeds along the x-axis and patients are organized on the y-axis. Warmer colors indicate a greater weight that the model puts on the corresponding time step.

AKI cohorts, which suggest that the groupwise classification is based heavily on these values. Interestingly, most of the top static features overlap with the RF, which boosts our confidence about their relevance for AKI prediction.

The temporal attention weights (bottom of Fig. 5) identify the intraoperative intervals that are deemed most relevant for AKI prediction. While the long and complicated surgical procedures make it difficult to pinpoint specific influences, we note that the attention is generally higher at the beginning and end of surgery, which coincide with the administration and withdrawal of anesthesia. This period can also reveal underlying issues, such as vascular stiffness. Hence, the results suggest that anesthetic strategies may be viable targets for optimization and that physiologic responses should be monitored closely during this time, as they may help to mitigate AKI.

V. Conclusion

We have presented two machine learning models to predict postoperative AKI: a random forest and a CNN-LSTM hybrid. Both models performed significantly better than our baseline methods, which are also commonly used in the AKI literature. While the random forest achieves the highest AUPRC, the CNN-LSTM goes one step beyond conventional frameworks by integrating intraoperative data and generating interpretable patient-specific feature weights via a dual attention mechanism. We also report multiple performance measures, which presents a more complete picture of model generalizability than the single AUROC metric reported in prior work. Lastly, our frameworks make minimal assumptions about the distribution of the data, thus allowing them to generalize in a cross-validated setting. From a clinical perspective, our models provide valuable benchmarks for AKI prediction and pave the way for future exploration on richer multimodal datasets.

References

[1] C. V. Thakar, "Perioperative acute kidney injury," *Advances in chronic kidney disease*, vol. 20, no. 1, pp. 67–75, 2013.

[2] F. Kork, F. Balzer, C. D. Spies, K.-D. Wernecke, A. A. Ginde, and et al., "Minor postoperative increases of creatinine are associated with higher mortality and longer hospital length of stay in surgical patients," *Anesthesiology*, vol. 123, no. 6, pp. 1301–1311, 2015.

[3] M. E. O'Connor, R. W. Hewson, C. J. Kirwan, G. L. Ackland, R. M. Pearse, and J. R. Prowle, "Acute kidney injury and mortality 1 year after major non-cardiac surgery," *BJS (British Journal of Surgery)*, vol. 104, no. 7, pp. 868–876, 2017.

[4] D. Ponce, C. d. P. F. Zorzenon, N. Y. Santos, and A. L. Balbi, "Early nephrology consultation can have an impact on outcome of acute kidney injury patients," *Nephrology Dialysis Transplantation*, vol. 26, no. 10, pp. 3202–3206, 2011.

[5] M. Küllmar, C. Massoth, M. Ostermann, S. Campos, N. Grau Novellas, G. Thomson, and et al., "Biomarker-guided implementation of the kdigo guidelines to reduce the occurrence of acute kidney injury in patients after cardiac surgery (prevaki-multicentre): protocol for a multicentre, observational study followed by randomised controlled feasibility trial," *BMJ Open*, vol. 10, no. 4, p. e034201, 2020.

[6] P.-Y. Tseng, Y.-T. Chen, C.-H. Wang, K.-M. Chiu, and et al., "Prediction of the development of acute kidney injury following cardiac surgery by machine learning," *Critical Care*, vol. 24, no. 1, pp. 1–13, 2020.

[7] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.

[8] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: point estimates and confidence intervals," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 451–466.

[9] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[10] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[11] H.-C. Lee and C.-W. Jung, "Vital recorder—a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[12] CKD-MBD Work Group *et al.*, "Kdigo clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease-mineral and bone disorder (ckd-mbd)." *Kidney international. Supplement*, no. 113, p. S1, 2009.

[13] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[14] C. Olah, "Understanding lstm networks," 2015.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] Y. Ma and G. Guo, *Support vector machines applications*. Springer, 2014, vol. 649.

[17] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[18] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize f1 score," *arXiv preprint ArXiv:1402.1892*, vol. 14, 2014.