Journal Pre-proof

Convolutional Neural Network Model for ICU Acute Kidney Injury Prediction

Sidney Le, BA, Angier Allen, BA, Jacob Calvert, MSc, Paul M. Palevsky, MD, Gregory Braden, MD, Sharad Patel, MD, Emily Pellegrini, MEng, Abigail Green-Saxena, PhD, Jana Hoffman, PhD, Ritankar Das, MSc

PII: S2468-0249(21)00102-9

DOI: https://doi.org/10.1016/j.ekir.2021.02.031

Reference: EKIR 1355

To appear in: *Kidney International Reports*

Received Date: 11 August 2020

Revised Date: 4 February 2021

Accepted Date: 15 February 2021

Please cite this article as: Le S, Allen A, Calvert J, Palevsky PM, Braden G, Patel S, Pellegrini E, Green-Saxena A, Hoffman J, Das R, Convolutional Neural Network Model for ICU Acute Kidney Injury Prediction, *Kidney International Reports* (2021), doi: https://doi.org/10.1016/j.ekir.2021.02.031.

Convolutional Neural Network Model for ICU Acute Kidney Injury Prediction

AUTHORS

Sidney Le BA[1+], Angier Allen BA[1+], Jacob Calvert MSc[1], Paul M. Palevsky MD[2], Gregory Braden MD[3], Sharad Patel MD[4], Emily Pellegrini MEng[1], Abigail Green-Saxena PhD[1*], Jana Hoffman PhD[1], Ritankar Das MSc[1]

[1] Dascena, Inc., Houston, Texas, United States
[2] VA Pittsburgh Healthcare System and University of Pittsburgh, Pittsburgh, Pennsylvania, United States
[3] Baystate Medical Center, Springfield, Massachusetts, United States
[4] Department of Critical Care Medicine, Cooper University Health Care

+Sidney Le and Angier Allen contributed equally to this work

* Corresponding author
Email: abigail@dascena.com
12333 Sowden Rd., Ste B, PMB 65148
Houston, Texas 77080-2059
(510) 826 - 9508

Support: This work was supported by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) [grant ID: 1R43AA02767401]

Running Headline: Development and validation of AKI prediction model

Keywords: acute kidney injury, machine learning, prediction, Electronic Health Record data, serum creatinine, convolutional neural net

ABSTRACT

Introduction: Acute kidney injury (AKI) is common among hospitalized patients and has a significant impact on morbidity and mortality. While early prediction of AKI has the potential to reduce adverse patient outcomes, it remains a difficult condition to predict and diagnose. The purpose of this study was to evaluate the ability of a machine learning algorithm to predict for AKI KDIGO Stage 2 or 3 up to 48 hours in advance of onset using convolutional neural networks (CNN) and patient Electronic Health Record (EHR) data.

Methods: A CNN prediction system was developed to use EHR data gathered during patients' stays to predict AKI up to 48 hours prior to onset. 12,347 patient encounters were retrospectively analyzed from the Medical Information Mart for Intensive Care III (MIMIC-III) database. **Comparators:** XGBoost AKI prediction model and the Sequential Organ Failure Assessment (SOFA) scoring system. **Outcomes:** AKI onset. **Analytical Approach:** The model was trained on routinely-collected patient EHR data. Measurements included Area Under the Receiver Operating Characteristic (AUROC) curve, positive predictive value (PPV), and a battery of additional performance metrics for advance prediction of AKI onset.

Results: On a hold-out test set, the algorithm attained an AUROC of 0.86 and PPV of 0.24, relative to a cohort AKI prevalence of 7.62%, for long-horizon AKI prediction at a 48-hour window prior to onset.

Conclusions: A CNN machine learning-based AKI prediction model outperforms XGBoost and the SOFA scoring system, demonstrating superior performance in predicting AKI 48 hours prior to onset, without reliance on serum creatinine measurements.

INTRODUCTION

Acute kidney injury (AKI) is a complex syndrome associated with large clinical and financial burdens. [1-12] Despite its prevalence in hospitalized patients[2,13] and reported incidence as high as 70% in the critically ill,[13,14] no treatment has been developed to effectively reverse injury to the kidney and restore kidney function.[1] The reasons for this failure have been attributed to delays in diagnosis and intervention,[2, 15-23] the complex nature of the AKI syndrome and the staging of its severity,[3, 21] and its multiple etiologies.[15,16]

Until recently, studies of incidence and outcomes of AKI have produced inconsistent results due to varying definitions of AKI.[24-26] The Risk, Injury, Failure, Loss, End-stage kidney disease (RIFLE) criteria,[27] followed by the Acute Kidney Injury Network (AKIN)[28] and most recently the Kidney Disease: Improving Global Outcomes (KDIGO) criteria[29, 30] have provided consensus on an AKI definition. KDIGO guidelines define acute kidney injury as an absolute increase of serum creatinine (SCr) of >0.3 mg/dL within 48 hours or a relative increase of >50% over no more than 7 days.[21, 29] Doubling of SCr at steady state reflects an approximate 50% decrease in kidney function as assessed by glomerular filtration rate (GFR).[31] Some studies have suggested that changes in SCr even smaller than 0.3 mg/dL within 48 hours are associated with significant increases in the risk of death, dialysis, and other morbidities,[6, 21, 32-38] and other studies are consistent with worsening outcomes with increasing AKI stage.[5, 24, 39-43] However, increases of serum creatinine are known to lag kidney injury by hours to days after the initial kidney insult, and therefore recognition of AKI is delayed by reliance on SCr measurements.[44,45]

Early AKI detection is critical to improving patient outcomes.[46-49] Given that the components necessary for defining and staging AKI are routinely available in the electronic health record (EHR),[3] a number of automated alerts have been developed to predict AKI events prior to onset. However, these alerts are generally triggered by detecting changes in SCr and/or urine output.[17] Because a range of kidney injury can exist before the loss of

kidney function can be estimated with these standard laboratory tests,[45,50] there is great interest in developing methods that could be used to detect AKI in patients at an earlier stage.[51-57] In this paper, we describe our methodology for the development of a convolutional neural network (CNN) prediction system that predicts AKI up to 48 hours prior to onset, using patient data extracted from the EHR. The CNN model does not require serum creatinine or urine output values.

METHODS

Description of data. This study uses data from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III version 1.3 dataset,[58] collected at Beth Israel Deaconess Medical Center in Boston, MA from 2001 to 2012. The MIMIC dataset offers a variety of encounter information from more than 40,000 unique patients and includes both structured (e.g. lab results) and unstructured (e.g. clinician notes) data. Due to differences in the storage of patient procedures information, we restrict our study to data collected from 2008 to 2012 using the MetaVision (iMDSoft) EHR system, and do not include data collected from 2001 to 2008 using the CareVue (Philips) system.[59] Because the collection of the MIMIC data did not affect patient safety and because all data were anonymized in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology have waived the requirement for patient consent.

From the MetaVision EHR MIMIC encounters, we selected for inclusion those stays involving adult patients (i.e. age 18 years or older) with at least one measurement of diastolic blood pressure, systolic blood pressure, temperature, respiratory rate, heart rate, SpO_2, and Glasgow Coma Scale. These measurements were selected because they are frequently available and easily collected at the patient bedside, even before clinical suspicion of AKI is present. These were the only direct variables used during training and testing of the algorithm; clinical notes vectorized with the Doc2Vec algorithm were also used as inputs to the CNN model. Serum creatinine was used as part of the KDIGO criteria, which served as the gold standard of AKI true positive patients, but was not

used as an input in testing. To facilitate the analysis of 48-hour advance prediction of AKI onset with a five-hour window of measurements upon which to base such a prediction, we required patient stay duration to be at least 53 hours in length. For convenience and with minimal restriction, we required that patient encounters lasted no more than 1000 hours. To train and test the algorithm on the broadest possible patient sample, no further inclusion or exclusion criteria were applied. Patients with prevalent AKI, with chronic kidney disease, or who receive dialysis were therefore included. Inclusion criteria are listed in **Figure 1** for 24 and 48 hour prediction windows, and the demographic characteristics of encounters meeting the inclusion criteria are reported in **Table 1**.

Overview of preprocessing, training, and testing. MIMIC-III ICU encounter data was gathered in the following way: encounters from the Metavision database in MIMIC-III were required to be at least 18 years of age, and had to include at least one measurement for at least one of the required input features. For each prediction offset *T*, the encounters were filtered such that each encounter was between $5+T$ hours and 1000 hours. $5+T$ hours was required to account for the offset, and to give the model the required 5 hours of measurements used for prediction. For each prediction offset *T*, positive examples' measurements were taken from between $5+T$ and *T* hours before onset to use for prediction, while negative examples' measurements were taken from random 5-hour windows in the patients' stays. Onset was defined as the first time that the relevant KDIGO criteria were met during the patient stay. Patient encounters satisfying the inclusion criteria were immediately allocated to training and testing sets. Roughly 90% and 10% of all encounters were randomly allocated to the training and testing sets, respectively, stratifying by positive and negative class to ensure equal representation of classes in both sets. We binned the data by the hour, imputed missing measurements, and standardized measurements on a variable-by-variable basis. AKI was defined according to KDIGO Stage 2 or KDIGO Stage 3 criteria and positive cases were identified as those patients reaching KDIGO Stage 2 or Stage 3 during the encounter. KDIGO Stage 2 or Stage 3 classifications were determined for each encounter, along with the corresponding times of KDIGO "onset" where appropriate. Stage 2 AKI is defined in the KDIGO staging system as an increase in SCr to more than 200% to 300% (>2- to 3-fold) from baseline or urine output <0.5

ml/kg per hour for more than 12 hours.[29] Stage 3 AKI is defined as an increase in SCr to more than 300% (>3-fold) from baseline, or ≥ 4.0 mg/dl (≥ 354 mmol/l), or kidney replacement therapy (KRT), or a decrease in estimated glomerular filtration rate (eGFR) to < 35 ml/min per 1.73m^2 (if <18 years of age), or urine output < 0.5 mL/kg/hr for ≥ 24 hours or anuria for ≥ 12 hours.[29] In both cases, the smaller of either the Modification of Diet in Renal Disease (MDRD)[27] serum creatinine estimate based on KDIGO 2012 guidelines or the 20th percentile of observed creatinine measurements was used for the baseline creatinine measurement in each patient encounter. Any missing features required for measurement, including missing urine or serum creatinine measures, made a contribution of 0 to the total KDIGO score.

A Doc2Vec embedding network was created to vectorize clinical text data. The Doc2Vec algorithm works by creating vectors for the most common words in all the documents, as well as separate vectors for each document. These vectors are trained by selecting a window of words in each document; the corresponding vectors for these words, in addition to the vector for the document that the text came from, to predict the next word in the sequence. The resulting document vectors are used as inputs, while the word vectors are discarded. The embedding network was prepared on a large collection of mid-stay clinical notes, ranging from primary complaint to radiology notes, including everything up to, but not including, the discharge summary, from encounters allocated to the training set. The network embedded texts into 250-dimensional numeric vectors, which served as inputs to the classifiers, alongside the structured data associated with the stays. Any notes dated after the onset of AKI were not used as inputs for the model to ensure the model is using only data found at or before prediction time.

Training data were passed to a convolutional neural network (CNN) structure, with hyperparameters optimized on the training set using the Python-based optimization package Talos. Tuned hyperparameters include learning rate, batch size, optimization loss, L1 and L2 regularization coefficients, and size of dense layers in the model. A CNN was chosen instead of a recurrent neural network (RNN) as they are faster to train and have fewer parameters.[60] In addition, the window of time from which the structured data was gathered for prediction was

relatively short (5 hours). CNN modeling techniques have been shown to outperform RNN modeling techniques with improved generalizability when applied to speech recognition tasks.[61] After the end of training on each fold, network performance was evaluated using the hold-out test set. Results were reported as the average test set performance across cross-validation folds.

Structured data preprocessing. Structured data were binned by the hour, with multiple intra-hour measurements of the same variable replaced by their average. Missing measurements were handled separately for training and testing sets using last observation carried forward imputation. Any remaining missing values were filled in using the measurement median over the training data. Quantitative data and document vectors were then standardized using the training data such that each feature has mean zero and variance one.

Document vector encoding network and unstructured data preprocessing. To facilitate the use of unstructured text data alongside the structured inputs, we trained a Doc2Vec[62] embedding network with 250 nodes, trained on 238,468 mid-stay clinical notes. Document vectors were produced for the text data available from each encounter, using 125 epochs of the Doc2Vec algorithm -- to better ensure the stability of inferred document vectors -- and an initial learning rate of 0.01. The choice of number of epochs and learning rate was found through experimentation. Clinical notes dated after AKI onset were excluded from inputs when training and testing the CNN.

Training of neural network classifier. We constructed a classifier to predict the probability of the presence of AKI at a given offset time from prediction using the Python deep learning library, Keras, that uses variants of multi-channel, multi-headed attention together with convolutions to extract information from the quantitative time series data. A separate network for handling the document vector produced by the Doc2Vec network was combined downstream through concatenation in a fully-connected output layer. This allows the model to incorporate information from both the time series data in the electronic health records, as well as the qualitative information found in the clinical notes. Model parameters were optimized using the Nadam optimizer[63] as

implemented in the Keras library with learning rate of 0.0009 and binary cross-entropy loss. A diagram of this neural network architecture is available as **Supplementary Figure S1**. Due to the low prevalence of AKI in the data, random oversampling was performed to artificially inflate the positive population. This was done by picking examples from the positive class at random with replacement until the number of positive examples matched the number of negative examples.

To fit the weights of the network with 10-fold cross-validation, we split the training data into 10 subsets of roughly equal size, and iteratively used 9 subsets for intra-fold training and the final subset for intra-fold testing. Model parameters were fit over the course of 50 epochs on the 9 intra-fold training subsets, with evaluation on the final subset. For each iterate, we obtained a receiver operating characteristic (ROC) curve, as well as a battery of performance metrics. We then randomly reset the model parameters before performing another iterate. From cross-validation, we obtained an average ROC curve and average performance metrics, along with standard deviation for the performance metrics. These results are presented in comparison with an XGBoost[64] classifier and the Sequential Organ Failure Assessment (SOFA) score,[65] which has been shown to independently predict AKI outcomes.[66-68] and therefore serves as a validated comparison measure for AKI prediction. SOFA was computed using all organ systems; any missing inputs required for computation contributed zero points to the total SOFA score. The XGBoost classifier was trained on the same processed training sets -- 5-hour windows of quantitative, clinical EHR data -- and evaluated on the same testing set. The time series data was turned into a list of the binned measurements at the different hours and given to XGBoost as input, requiring no additional feature engineering. Document vectors were not given as input for XGBoost. XGBoost hyperparameters were tuned using a cross-validated grid search on the training data. Hyperparameters were optimized using grid-search over the hyperparameters "*gamma*", which controls how often the trees are split, and "*colsample_bytree*", which controls the number of features randomly selected for inputs when constructing each tree.

RESULTS

The demographic characteristics associated with MIMIC III ICU encounters meeting the inclusion criteria of **Figure 1** are provided in **Table 1**. The study population was 53.29% male, with few (4.65%) patients younger than 30 years of age and a substantial percentage of patients aged 70 years or more (39.39%). More than half (62.78%) of patients had stays lasting between 3 and 5 days, with a substantial percentage of patients experiencing stays of 12 days or longer (11.26%). The overall mortality rate was 39.39%, with 7.62% of encounters meeting the criteria for KDIGO Stage 2 or Stage 3 at some point in the stay, and 20.6% of stays meeting some stage of the KDIGO criteria at any point during the stay.

Performance is evaluated by predicting once in each encounter using 5 hours of data. This data is taken either from a random portion of the stay, for negative examples, or from the specified model's offset for positive examples. The results from 10-fold cross-validation on the 90% training set are reported in **Tables 2 and 3, for 48 and 24 hr predictions, respectively.** Test performance is reported for the best-performing model, selected by cross-validation of the train data. The CNN model with the use of the Doc2Vec embeddings of encounter text data outperformed the XGBoost comparator model and the SOFA score for advance prediction of KDIGO Stage 2 or Stage 3 onset. We note that, in order to provide non-summative performance metrics (i.e., the metrics other than area under the receiver operating characteristic (AUROC) curve), we selected an operating point for each model or score which provided a sensitivity nearest 0.80. The CNN model performed better (AUROC of 0.86 for 24 and 48 hr predictions) when text data were made available through Doc2Vec than when these data were unavailable (AUROC of 0.77 and 0.76 for 24 and 48 hr predictions, respectively). In addition, the quality of prediction was higher for KDIGO Stage 2 or Stage 3 onset, as compared with the prediction of onset for any of KDIGO Stages 1-3. For corresponding CNN and XGBoost results without oversampling of the minority class, see **Supplementary Table S1**. Permutation feature importance methods were implemented to provide information on the relative importance of each input variable. A precision-recall curve comparison between the CNN model, the XGBoost model, and the SOFA score is presented in **Supplementary Figure S2.**

The CNN model averaged a positive predictive value (PPV) of 0.24 over cross-validation folds for the 48-hour prediction of KDIGO Stages 2 and 3, compared to average PPVs of 0.09 and 0.13 for XGBoost and the SOFA score, respectively (**Table 2**). The advantage of the CNN mostly vanished (PPV of 0.16) in the absence of text data through Doc2Vec input. The average PPV was highest when the CNN classifier was given access to Doc2Vec input and tasked with 48-hour prediction of KDIGO Stages 1-3 (PPV of 0.31). Relative to the 7.62% prevalence of KDIGO Stages 2 and 3, positive predictions made by the CNN model enriched for KDIGO Stage 2 or 3 encounters by a factor of 4.80, whereas XGBoost and the SOFA score enriched these encounters by factors of 2.50 and 2.11, respectively.

The ROC curve comparison of 48-hour prediction on the 10% hold-out test set is shown in **Figure 2**. The CNN model, which was provided text data through Doc2Vec input, performed substantially better than the XGBoost model and the SOFA score. The XGBoost model and SOFA had similar performance on the test set.

DISCUSSION

These experiments demonstrate that a convolutional neural network can predict AKI up to 48 hours in advance of KDIGO Stage 2 or Stage 3 AKI onset, with AUROC performance superior to that of an XGBoost classifier and the SOFA scoring system (**Table 2, Figure 2**). Unlike other diseases for which multiple severity scores exist, AKI represents a group of syndromes that are loosely connected by the characteristic rapid drop in eGFR seen in AKI patients.[69] With over 30 definitions of AKI,[70] attempts at a uniform definition for acute kidney injury have included the RIFLE classification[27] followed by the Acute Kidney Injury Network (AKIN)[28] and most recently the KDIGO criteria.[29, 30] The absence of a consistent uniform definition may explain the current lack of an AKI-specific risk score that serves as a standard of care. To provide context for the performance of their models, prior studies focusing on the development of AKI prediction models have either used the biomarker serum neutrophil gelatinase associated lipocalin (sNGAL) as a comparator,[71] compared their model

to other ML models[72] or not included a standard of care comparator.[73,74] In the current study, we compare two ML models as well as provide the SOFA score as a comparator. Although the SOFA score was not developed for the purpose of long-horizon AKI prediction, because of the ubiquity of the SOFA score, and previous usage in AKI outcome prediction, it serves as a validated comparator for our current approach.[66-68] The XGBoost comparator is similarly important, primarily due to its broad and successful use in applications to other clinical prediction tasks (e.g., the 2019 Physionet Computing in Cardiology Challenge[75]).

The superiority of the CNN classifier to the XGBoost classifier and the commonly-used SOFA score is evidenced by key performance metrics, such as AUROC and PPV (**Table 2**). The PPV performance improvement is of particular importance. Romero-Brufau *et al.* have argued that AUROC performance may be misleading for clinicians interested in evaluating the clinical impact of a diagnostic tool, as AUROC does not incorporate information about the prevalence of a condition.[76] In fact, for the same reason, AUROC is useful for comparing the performance of tools retrospectively validated on different datasets. This concern regarding PPV and prevalence is relevant to our study, as we found that the prevalence of KDIGO Stages 2 or 3 is roughly 7.6% in the cohort, an estimate consistent with prior epidemiologic studies.[77] The AUROC is a summative metric which may include ranges of operating points which are irrelevant to a given task, whereas PPV can be focused on a clinically relevant operating point. To produce the metrics in **Table 2**, we chose operating points for the CNN and comparators which fixed their sensitivities near 0.80.

Beyond the text data input through Doc2Vec, CNN predictions were made using only age and 7 routinely collected patient measurements (diastolic blood pressure, systolic blood pressure, temperature, respiratory rate, heart rate, SpO_2, and Glasgow Coma Scale) as inputs. Although this study was restricted to MetaVision (iMDSoft) EHR system for technical reasons, the use of these widely available inputs supports that the model could be generalized to broad clinical practice. Importantly, the CNN model did not rely on SCr to make predictions, distinguishing it from other AKI prediction tools. Creatinine levels can take hours or days to rise to AKI thresholds as defined in the KDIGO staging system[78]; therefore, changes in SCr may reflect pre-existing

kidney damage. An AKI prediction tool which does not depend on SCr measurements may better afford clinicians the opportunity to intervene early, to prevent AKI development or progression, or to limit further kidney damage. Additionally, using only commonly collected variables in the EHR for AKI prediction allows automatic screening of a general patient population for impending AKI without requiring specialized evaluation.

This study contributes to the growing body of retrospective machine learning (ML) literature for the prediction of AKI.[79] Chiofolo et al. (2019) developed a model for AKI prediction and surveillance in ICU patients at a 6-hr prediction window with an AUROC of 0.88.[73] Fletchet et al. (2017) developed the AKIpredictor, a prognostic calculator for prediction of AKI in ICU patients during the first week of stay.[71] Their KDIGO Stage 2 and 3 model produced AUROCs between 0.77 and 0.84. The AUROC of 0.84 corresponds to a prediction of KDIGO Stage 2 and 3 after gathering 24 hours of data. As a point of comparison, the CNN model used only 5 hours of data before making a prediction. Recent work by Tomasev *et al.* pursued a deep learning approach for continuous risk prediction of deterioration in acute kidney injury patients, and evaluated their tool on a Veteran's Health Administration dataset of 703,782 adult patients.[72] Algorithm performance at a 48-hour prediction window corresponded to a sensitivity of 55.8% and a specificity of 82.7%.[72] This performance is reported to be "in range" required for regulatory approval.[80] While these studies make important contributions to the domain of AKI research, they depend on the use of SCr to make predictions, which is a lagging marker of kidney function. In contrast, the CNN described in this work does not rely on SCr to make predictions of AKI onset, which allows for both longer lead times and improved predictive performance, as well as for generating predictions on patients who may not yet be clinically suspected of having AKI and have not yet had SCr measures drawn. The CNN also offers improvement in performance as compared to our prior work, [81] which utilized the machine learning method of gradient boosted trees to predict AKI prior to onset and included SCr as a model input. In comparison, results from our current work suggest that AKI predictions can be made with a

more robust machine learning architecture, without reliance on SCr, while achieving stronger predictive performance.

While the CNN described in this study offers substantial lead time in AKI identification (up to 48 hours), and offers improved predictive performance over our previous work,[81] it still requires prospective validation. Additionally, we cannot determine from this retrospective study what impact the algorithm might have on clinicians and their provision of care in clinical settings, nor provide an analysis of model evaluation and its prediction performance in time. While the CNN model performance was superior to that of SOFA and XGBoost, improvements in PPV achieved by the CNN compared to XGBoost or SOFA are less pronounced without the use of clinical notes. Algorithm performance is assessed only on US patients older than age 18, with stays in the ICU, which limits the generalizability of our results to other patient populations and levels of care. While the majority of the negative class patients had a serum creatinine measurement at some point in the ICU stay, it is possible that inclusion of patients missing urine measures in the negative class led to the misclassification of some patients in our dataset. It is also possible that misclassifications could have occurred for some patients in the data set due to inclusion of patients with a previous diagnosis of CKD and/or who received dialysis. Due to a lack of a standard-of-care AKI score, we used the SOFA score and the XGBoost model to provide context for our model performance. While the SOFA score has been used in AKI outcome prediction studies,[66-68] it was not developed for the purpose of long-horizon AKI prediction. Further, while the XGBoost comparator was included due to its use in other clinical prediction tasks,[75] it does not serve as a standard-of-care for AKI predictions. Lastly, because there have been several proposed consensus definitions for AKI, the algorithm we described may have different results when compared against non-KDIGO definitions, or in settings which utilize a different standard in their diagnostic procedures.

CONCLUSION

A convolutional neural network for AKI prediction outperforms XGBoost and the traditional SOFA scoring system, demonstrating superior performance in predicting acute kidney injury up to 48 hours prior to onset, without reliance on measurements of changes in serum creatinine. Although the use of clinical text data through a Doc2Vec network substantially strengthens CNN prediction performance, the CNN demonstrated superior performance over both XGBoost and SOFA even when clinical notes were not included as model inputs, supporting the use of CNN models for the task of AKI prediction. Such a tool may improve prediction and early detection of AKI in clinical settings, thereby allowing for earlier intervention.

Disclosures

Author's Contributions:

R.D., S.L., and A.A. conceived and designed this study; S.L. and A.A. performed the modeling and statistical analysis; all authors contributed to acquisition, analysis, or interpretation of data; S.L., J.H., A.S., and E.P. drafted the article; all authors revised the article for important intellectual content; and R.D. obtained funding.

Support: This work was supported by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) [grant ID: 1R43AA02767401]

Financial Disclosures: All authors who have affiliations listed with Dascena (Houston, Texas, USA) are employees or contractors of Dascena.

Data Sharing Plan: The data that support the findings of this study are publicly available from http://www.nature.com/articles/sdata201635.

SUPPLEMENTARY MATERIAL

1. Supplementary Figure S1 (PDF): Schematic diagram of the neural network architecture.
2. Supplementary Table S1 (PDF): Model performance metric results from 10-fold cross-validation.

3. Supplementary Figure S2 (PDF): Precision-recall curve comparison of model prediction performance.

Supplementary information is available at KI Report's website.

REFERENCES

1. Kashani K, Ronco C. Acute Kidney Injury Electronic Alert for Nephrologist: Reactive versus Proactive? Blood Purif 2016;42:323–328. doi: 10.1159/000450722
2. Al-Jaghbeer M, Dealmeida D, Bilderback A, Ambrosino R, Kellum JA. Clinical Decision Support for In- Hospital AKI. Journal of the American Society of Nephrology. 2017 Nov 2:ASN-2017070765.
3. Hoste EAJ, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, Edipidis K, Forni LG, Gomersall CD, Govil D, Honoré PM, Joannes-Boyau O, Joannidis M, Korhonen A-M, Lavrentieva A, Mehta RL, Palevsky P, Roessler E, Ronco C, Uchino S, Vazquez JA, Vidal Andrade E, Webb S, Kellum JA. Epidemiology of acute kidney injury in critically ill patients: The multinational AKI-EPI study. Intensive Care Med 2015; 41: 1411–1423.
4. Wang HE, Muntner P, Chertow GM, Warnock DG: Acute kidney injury and mortality in hospitalized patients. Am J Nephrol 2012;35: 349–355.
5. Uchino S, Bellomo R, Goldsmith D, Bates S, Ronco C: An assessment of the RIFLE criteria for acute renal failure in hospitalized patients. Crit Care Med 2006;34: 1913–1917.
6. Chertow GM, Burdick E, Honour M, Bonventre JV, Bates DW: Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. J Am Soc Nephrol 2005; 16: 3365–3370.
7. Kellum JA, Sileanu FE, Murugan R, Lucko N, Shaw AD, Clermont G: Classifying AKI by urine output versus serum creatinine level. J Am Soc Nephrol 2015; 26: 2231–2238.
8. Hoste EAJ, Kellum JA, Selby NM, Zarbock A, Palevsky PM, Bagshaw SM, Goldstein SL, Cerdá J, Chawla LS. Global epidemiology and outcomes of acute kidney injury. Nat Rev Nephrol. 2018 Oct;14(10):607-625. doi: 10.1038/s41581-018-0052-0. Review. PubMed PMID: 30135570.
9. Freda BJ, Knee AB, Braden GL, Visintainer PF, Thakar CV. Effect of Transient and Sustained Acute Kidney Injury on Readmissions in Acute Decompensated Heart Failure. Am J Cardiol. 2017 Jun 1;119(11):1809-1814.
10. Palevsky PM, Zhang JH, O'Connor TZ, Chertow GM, Crowley ST, Choudhury D, Finkel K, Kellum JA, Paganini E, Schein RMH, Smith MW, Swanson KM, Thompson BT, Vijayan A, Watnick S, Star RA, Peduzzi P. Intensity of renal support critically ill patients with acute kidney injury. *New England Journal of Medicine*. 2008; 359: 7-20.
11. Kellum JA, Chawla LS, Keener C, Singbartl K, Palevsky PM, Pike FL, Yealy DM, Huang DT, Angus DC. The effects of alternative resuscitation strategies in acute kidney injury patients with septic shock. American Journal of Respiratory and Critical Care Medicine. 2016; 193: 281-287.
12. Khalil P, Murty P, Palevsky PM. The patient with acute kidney injury. Prim Care. 2008 Jun;35(2):239-64, vi. doi: 10.1016/j.pop.2008.01.003. Review. PubMed PMID: 18486715.
13. Forni LG, Dawes T, Sinclair H, et al. Identifying the patient at risk of acute kidney injury a predictive scoring system for the development of acute kidney injury in acute medical patients. Nephron Clinical Practice. 2013;123(3-4): 143-150.
14. Uchino S, Kellum JA, Bellomo R, et al: Beginning and Ending Supportive Therapy for the Kidney (BEST Kidney) Investigators: Acute renal failure in critically ill patients: a multinational, multicenter study. JAMA 2005;294: 813–818.
15. Endre JH, Pickering JW. Acute kidney injury clinical trial design: old problems, new strategies. Pediatr Nephrol. 2013;28: 207–217. doi: 10.1007/s00467-012-2171-3
16. Pickering JW, Ralib AM, Nejat M, Endre ZH. New considerations in the design of clinical trials of acute kidney injury. Clin Invest 2011;1: 637–650.

17. Lachance P, Villeneuve PM, Rewa OG, et al. Association between e-alert implementation for detection of acute kidney injury and outcomes: a systematic review. Nephrol Dial Transplant. 2017;32(2):265-272.
18. Kolhe NV, Staples D, Reilly T et al. Impact of compliance with a care bundle on acute kidney injury outcomes: a prospective observational study. PLoS One 2015; 10: e0132279
19. Terrell KM, Perkins AJ, Hui SL et al. Computerized decision support for medication dosing in renal insufficiency: a randomized, controlled trial. Ann Emerg Med 2010; 56: 623–629
20. Colpaert K, Hoste EA, Steurbaut K et al. Impact of real-time electronic alerting of acute kidney injury on therapeutic intervention and progression of RIFLE class. Crit Care Med 2012; 40: 1164–1170
21. Wilson FP, Shashaty M, Testani J et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. Lancet 2015; 385: 1966–1974
22. Thomas ME, Sitch A, Baharani J et al. Earlier intervention for acute kidney injury: evaluation of an outreach service and a long-term follow-up. Nephrol Dial Transplant 2015; 30: 239–244
23. Jo S-K, Rosner MH, Okusa MD. Pharmacologic treatment of acute kidney injury: why drugs haven't worked and what is on the horizon. Clin J Am Soc Nephrol 2007;2: 356–365.
24. Porter CJ, Juurlink I, Bisset LH, Bavakunji R, Mehta RL, Devonald MA. A real-time electronic alert to improve detection of acute kidney injury in a large teaching hospital. Nephrol Dial Transplant. 2014 Oct;29(10):1888-93. doi: 10.1093/ndt/gfu082.
25. Waikar SS, Curhan GC, Wald R et al. Declining mortality in patients with acute renal failure, 1988 to 2002. J Am Soc Nephrol 2006;17: 1143–1150.
26. Xue JL, Daniels F, Star RA et al. Incidence and mortality of acute renal failure in Medicare beneficiaries, 1992 to 2001. J Am Soc Nephrol 2006; 17: 1135–1142
27. Bellomo R, Ronco C, Kellum JA et al. Acute renal failure—definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care 2004; 8: R204–R212.
28. Mehta RL, Kellum JA, Shah SV et al. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 2007; 11: R31.
29. Kidney Disease Improving Global Outcomes. Clinical Practice Guideline for Acute Kidney Injury. Kidney Int 2012; 2:1-138.
30. Palevsky PM, Liu KD, Brophy PD, Chawla LS, Parikh CR, Thakar CV, Tolwani AJ, Waikar SS, Weisbord, SD. KDOQI US Commentary on the 2012 KDIGO Clinical Practice Guideline for Acute Kidney Injury. American Journal of Kidney Diseases. 2013; 61: 649-672.
31. Weisenthal SJ, Quill C, Farooq S, Kautz H, Zand MS. Predicting acute kidney injury at hospital re-entry using high-dimensional electronic health record data. PLoS ONE 2018;13(11): e0204920.
32. Joannidis M, Metnitz B, Bauer P, et al. Acute kidney injury in critically ill patients classified by AKIN versus RIFLE using the SAPS 3 database. Intensive Care Med. 2009; 35:1692–702.
33. Kolli H, Rajagopalam S, Patel N, et al. Mild acute kidney injury is associated with increased mortality after cardiac surgery in patients with eGFR <60 mL/min/1·73 m(2). Ren Fail. 2010; 32:1066–72.
34. Wilson FP, Yang W, Feldman HI. Predictors of death and dialysis in severe AKI: the UPHS-AKI cohort. Clin J Am Soc Nephrol. 2013; 8:527–37.
35. Bihorac A, Delano MJ, Schold JD, et al. Incidence, clinical predictors, genomics, and outcome of acute kidney injury among trauma patients. Ann Surg. 2010; 252:158–65.
36. Bihorac A, Yavas S, Subbiah S, et al. Long-term risk of mortality and acute kidney injury during hospitalization after major surgery. Ann Surg. 2009; 249:851–58.
37. Garzotto F, Piccinni P, Cruz D, et al. RIFLE-based data collection/management system applied to a prospective cohort multicenter Italian study on the epidemiology of acute kidney injury in the intensive care unit. Blood Purif. 2011;31:159–71.
38. Newsome BB, Warnock DG, McClellan WM, et al. Long-term risk of mortality and end-stage renal disease among the elderly after small increases in serum creatinine level during hospitalization for acute myocardial infarction. Arch Intern Med. 2008; 168:609–16.

39. Coca SG, Yusuf B, Shlipak MG et al. Long-term risk of mortality and other adverse outcomes after acute kidney injury: a systematic review and meta-analysis. Am J Kidney Dis 2009; 53: 961–973.
40. Lafrance JP, Miller DR. Acute kidney injury associates with increased long-term mortality. J Am Soc Nephrol 2010; 21: 345–352.
41. Ricci Z, Cruz D, Ronco C. The RIFLE criteria and mortality in acute kidney injury: a systematic review. Kidney Int 2008; 73: 538–546.
42. Uchino S, Kellum JA, Bellomo R et al. Acute renal failure in critically ill patients: a multinational, multicenter study. JAMA 2005; 294: 813–818.
43. Ali T, Khan I, Simpson W et al. Incidence and outcomes in acute kidney injury: a comprehensive population-based study. J Am Soc Nephrol 2007; 18: 1292–1298.
44. Ostermann M, Joannidis M. Acute kidney injury 2016: diagnosis and diagnostic workup. Critical care. 2016 Dec 1;20(1):299.
45. Makris K. The role of the clinical laboratory in the detection and monitoring of acute kidney injury. Journal of Laboratory and Precision Medicine. 2018 Oct 8;3.
46. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017; 24(6):1052–1061.
47. Park S, Ha Baek S, Ahn S, et al. Impact of Electronic Acute Kidney Injury (AKI) Alerts With Automated Nephrologist Consultation on Detection and Severity of AKI: A Quality Improvement Study. Am J Kidney Dis .2018 Jan; 71(1): 9-19. doi: 10.1053/j.ajkd.2017.06.008
48. de Virgilio C, Kim DY. Transient acute kidney injury in the postoperative period: it is time to pay closer attention. JAMA Surg. 2016;151(5):450-451.
49. Soares DM, Pessanha JF, Sharma A, Brocca A, Ronco C. Delayed nephrology consultation and high mortality on acute kidney injury: a meta-analysis. Blood Purif. 2016;43(1-3):57-67.
50. Thomas ME, Blaine C, Dawnay A, Devonald MA, Ftouh S, Laing C, et al. The definition of acute kidney injury and its use in practice. Kidney Int. 2015;87:62–73.
51. Hodgson LE, Dimitrov BD, Roderick PJ, et al. Predicting AKI in emergency admissions: an external validation study of the acute kidney injury prediction score (APS). BMJ Open 2017;7:e013511.
52. Braitman LE, Davidoff F. Predicting clinical states in individual patients. Ann Intern Med 1996;125:406–12.
53. Feinstein AR. “Clinical Judgment” revisited: the distraction of quantitative models. Ann Intern Med 1994;120:799–805.
54. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201–10.
55. Christensen E. Prognostic models including the Child-Pugh, MELD and Mayo risk scores--where are we and where should we go? J Hepatol 2004;41:344–50.
56. Kashani K, Rosner MH, Haase M, Lewington AJP, O'Donoghue DJ, Wilson FP, Nadim MK, Silver SA, Zarbock A, Ostermann M, Mehta RL, Kane-Gill SL, Ding X, Pickkers P, Bihorac A, Siew ED, Barreto EF, Macedo E, Kellum JA, Palevsky PM, Tolwani AJ, Ronco C, Juncos LA, Rewa OG, Bagshaw SM, Mottes TA, Koyner JL, Liu KD, Forni LG, Heung M, Wu VC. Quality Improvement Goals for Acute Kidney Injury. Clin J Am Soc Nephrol. 2019 Jun 7;14(6):941-953. doi: 10.2215/CJN.01250119. Epub 2019 May 17. PubMed PMID: 31101671; PubMed Central PMCID: PMC6556737.
57. Garcia S, Bhatt DL, Gallagher M, Jneid H, Kaufman J, Palevsky PM, Wu H, Weisbord SD. Strategies to Reduce Acute Kidney Injury and Improve Clinical Outcomes Following Percutaneous Coronary Intervention: A Subgroup Analysis of the PRESERVE Trial. JACC Cardiovasc Interv. 2018 Nov 26;11(22):2254-2261. doi: 10.1016/j.jcin.2018.07.044. PubMed PMID: 30466822.
58. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635
59. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR medical informatics. 2016;4(3):e28.

60. Blohm M, Jagfeld G, Sood E, Yu X, Vu NT. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. arXiv preprint arXiv:1808.08744. 2018 Aug 27.
61. Oord AV, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. 2016 Sep 12.
62. Le Q, Mikolov T. Distributed representations of sentences and documents. International conference on machine learning 2014 Jan 27 (pp. 1188-1196)
63. Dozat, T. Incorporating Nesterov Momentum into Adam. ICLR Workshop, (1):2013–2016, 2016.
64. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794). ACM.
65. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive care medicine. 1996 Jul;22(7):707.
66. Hoste EAJ, Clermont G, Kersten A, et al. RIFLE criteria for acute kidney injury is associated with hospital mortality in critically ill patients: a cohort analysis. Crit Care 2006, 10: R73-82. 10.
67. De Mendonca A, Vincent JL, Suter PM, Moreno R, Dearden NM, Antonelli M, Takala J, Sprung C, Cantraine F. Acute renal failure in the ICU: risk factors and outcomes evaluated by the SOFA score. Intensive Care Med. 2000 July; 26:915-921.
68. Chang CH, Fan PC, Chang MY, Tian YC, Hung CC, Fang JT, Yang CW, Chen YC. Acute kidney injury enhances outcome prediction ability of sequential organ failure assessment score in critically ill patients. PloS one. 2014 Oct 3;9(10):e109649.
69. Kellum JA, Prowle JR. Paradigms of acute kidney injury in the intensive care setting. Nature Reviews Nephrology. 2018 Apr;14(4):217.
70. Kellum JA, Levin N, Bouman C, Lameire N. Developing a consensus classification system for acute renal failure. Current opinion in critical care. 2002 Dec 1;8(6):509-14.
71. Flechet M, Güiza F, Schetz M, et al. AKIpredictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. Intensive Care Med. 2017;43(6):764-773.
72. Tomasev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572:116–119.
73. Chiofolo C, Chbat N, Ghosh E, et al. Automated continuous acute kidney injury prediction and surveillance: a random forest model. Mayo Clin Proc. May 2019;94(5):783-792.
74. Simonov M, Ugwuowo U, Moreira E, Yamamoto Y, Biswas A, Martin M, Testani J, Wilson FP. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: A descriptive modeling study. PLoS medicine. 2019 Jul 15;16(7):e1002861.
75. Reyna M, Josef C, Jeter R, Shashikumar S, Westover M, Nemati S, Clifford G, Sharma A. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine. 2019 Oct 14.
76. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Critical Care. 2015 Dec;19(1):285.
77. Case J, Khan S, Khalid R, Khan A. Epidemiology of acute kidney injury in the intensive care unit. Crit Care Res Pract. 2013;2013:479730. doi: 10.1155/2013/479730. Epub 2013 Mar 21. PMID: 23573420; PMCID: PMC3618922.
78. Waikar SS, Bonventre JV. Creatinine kinetics and the definition of acute kidney injury. Journal of the American Society of Nephrology. 2009 Mar 1;20(3):672-9.
79. Palevsky PM. Electronic Alerts for Acute Kidney Injury. Am J Kidney Dis. 2018 Jan;71(1):1-2. doi: 10.1053/j.ajkd.2017.09.009. PubMed PMID: 29273153.
80. Kellum J, Bihorac A. Artificial intelligence to predict AKI: is it a breakthrough? Nature Reviews Nephrology 2019;15:663–664.

81. Mohamadlou H, Lynn-Palevsky A, Barton C, Chettipally U, Shieh L, Calvert J, Saber NR, Das R. Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data. Can J Kidney Health Dis. 2018 Jun 8;5:2054358118776326. doi: 10.1177/2054358118776326. PMID: 30094049; PMCID: PMC6080076.

TABLES

Table 1. Demographic characteristics of MIMIC III ICU encounters found in the 48hr dataset and meeting the inclusion criteria of Figure 1. We note that the determination of KDIGO positive or negative was made after the data preprocessing steps described in the Methods section.

Characteristic		Count	Percent
Gender	Female	3,186	46.71
	Male	3,186	53.29
Age (days): Median 65, IQR (53-77)	18-29	317	4.65
	30-39	307	4.50
	40-49	665	9.75
	50-59	1,246	18.27
	60-69	1,599	23.44
	70+	2,687	39.39
Length of Stay (days): Median 5, IQR (4-9)	< 3	43	0.63
	3-5	4,282	62.78
	6-8	1,200	17.59
	9-11	528	7.74
	12+	768	11.26
In-Hospital Death	Yes	1,747	25.61
	No	5,074	74.39
KDIGO 2/3	Positive	520	7.62
	Negative	6,301	92.38
KDIGO 1/2/3	Positive	1,410	20.67
	Negative	5,411	79.33

Table 2. Results from 10-fold cross-validation of predictions **48 hours prior to onset** on the MIMIC III data set. The convolutional neural network (CNN) model is compared with an XGBoost classifier, and the Sequential Organ Failure Assessment (SOFA) score. SOFA required no training and thus could be applied to the entire test set at once, hence no standard deviation is reported. Additional comparison is made to the CNN model without the use of the Doc2Vec network (i.e., without unstructured text data) and for the prediction of KDIGO criteria of any stage. Abbreviations: area under the receiver operating characteristic (AUROC) curve; diagnostic odds ratio (DOR); positive and negative likelihood ratios (LR+ and LR-, respectively); positive and negative predictive value (PPV and NPV, respectively); standard deviation (SD).

	CNN	XGBoost	SOFA	No Doc2Vec	Stage 1 Included	Stage 3 Only
AUROC mean (SD)	0.856 (0.034)	0.654 (0.011)	0.701	0.763 (0.035)	0.778 (0.037)	0.819 (0.036)
Sensitivity mean (SD)	0.804 (0.000)	0.798 (0.000)	0.798	0.805 (0.006)	0.806 (0.008)	0.806 (0.000)
Specificity mean (SD)	0.763 (0.057)	0.380 (0.006)	0.441	0.623 (0.064)	0.649 (0.074)	0.679 (0.079)
PPV mean (SD)	0.236 (0.039)	0.095 (0.001)	0.127	0.163 (0.022)	0.310 (0.044)	0.105 (0.023)
NPV mean (SD)	0.975 (0.002)	0.956 (0.001)	0.960	0.970 (0.003)	0.940 (0.006)	0.985 (0.002)
Accuracy mean (SD)	0.765 (0.052)	0.411 (0.005)	0.612	0.638 (0.056)	0.672 (0.062)	0.683 (0.076)
DOR mean (SD)	14.076 (3.779)	2.421 (0.059)	3.123	7.123 (1.899)	8.167 (2.425)	9.566 (3.410)
LR+ mean (SD)	3.558 (0.739)	1.287 (0.012)	1.429	2.191 (0.362)	2.389 (0.478)	2.658 (0.660)
LR- mean (SD)	0.258 (0.021)	0.532 (0.008)	0.458	0.316 (0.035)	0.301 (0.035)	0.288 (0.035)
F1 mean (SD)	0.361 (0.047)	0.169 (0.001)	0.214	0.270 (0.030)	0.444 (0.045)	0.184 (0.036)

Table 3. Results from 10-fold cross-validation of predictions **24 hours prior to onset** on the MIMIC III data set. The convolutional neural network (CNN) model is compared with an XGBoost classifier, and the Sequential Organ Failure Assessment (SOFA) score. SOFA required no training and thus could be applied to the entire test set at once, hence no standard deviation is reported. Additional comparison is made to the CNN model without the use of the Doc2Vec network (i.e., without unstructured text data) and for the prediction of KDIGO criteria of any stage. Abbreviations: area under the receiver operating characteristic (AUROC) curve; diagnostic odds ratio (DOR); positive and negative likelihood ratios (LR+ and LR-, respectively); positive and negative predictive value (PPV and NPV, respectively); standard deviation (SD).

	CNN	XGBoost	SOFA	No Doc2Vec	Stage 1 Included	Stage 3 Only
AUROC mean (SD)	0.863 (0.009)	0.729 (0.009)	0.727	0.769 (0.028)	0.834 (0.004)	0.867 (0.009)
Sensitivity mean (SD)	0.803 (0.000)	0.801 (0.000)	0.784	0.801 (0.003)	0.798 (0.005)	0.795 (0.000)
Specificity mean (SD)	0.772 (0.021)	0.463 (0.026)	0.537	0.585 (0.066)	0.716 (0.018)	0.785 (0.024)
PPV mean (SD)	0.221 (0.016)	0.111 (0.005)	0.151	0.153 (0.019)	0.359 (0.014)	0.131 (0.014)
NPV mean (SD)	0.978 (0.001)	0.964 (0.002)	0.961	0.968 (0.003)	0.944 (0.001)	0.988 (0.000)
Accuracy mean (SD)	0.773 (0.020)	0.489 (0.024)	0.684	0.602 (0.060)	0.728 (0.014)	0.784 (0.023)
DOR mean (SD)	13.905 (1.617)	3.484 (0.367)	4.200	5.861 (1.440)	10.030 (0.822)	14.396 (2.212)
LR+ mean (SD)	3.545 (0.319)	1.494 (0.073)	1.692	1.970 (0.292)	2.821 (0.178)	3.740 (0.452)
LR- mean (SD)	0.256 (0.007)	0.431 (0.024)	0.403	0.344 (0.038)	0.282 (0.007)	0.261 (0.008)
F1 mean (SD)	0.345 (0.019)	0.194 (0.007)	0.247	0.256 (0.027)	0.494 (0.013)	0.224 (0.020)

FIGURES LEGENDS

Figure 1. Inclusion diagram. Patients were required to be at least 18 years of age, and must have at least one measurement of at least one of the input features.

Figure 2. ROC curve comparison of prediction performance using a convolutional neural net (CNN) classifier, an XGBoost (XGB) classifier, and the SOFA score, 48 hours prior to AKI onset on the MIMIC III ICU hold out data set. AUROC, Area Under the Receiver Operating Characteristic curve; SOFA, Sequential Organ Failure Assessment score.