Predicting of Acute Kidney Injury using Electronic Health Records

by

Svetlana Maslenkova

Thesis submitted to the
Deanship of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the M.Sc. degree in
Machine Learning

Deprtement of Machine Learning
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor(s): Dr. Mohammad Yaqub
Assistant Professor, Department of Computer Vision,
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Internal Member: Dr. Muhammad Abdul-Mageed
Associate Professor,
Department of Natural Language Processing,
Department of Machine Learning,
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Acute Kidney Injury (AKI) affects more than 13 million people annually and increases the risk of death in patients[26]. The severity of AKI also contributes to the increase in associated costs of a patient's treatment. The early prediction of AKI could enable clinicians to focus on preventive treatment for at-risk patients. The adoption of Electronic Health Records (EHR) in medical institutions allows healthcare professionals to access patients' health more efficiently and develop personalized treatment trajectories, thereby improving healthcare quality. Therefore, the AKI risk prediction algorithm based on EHR could enable clinicians to devote more time directly to treating patients instead of reviewing tons of information related to patients' visits.

In this thesis, we used the publicly available EHR database MIMIC-IV v2.0 to develop an AKI risk prediction framework for patients admitted to Intensive Care Units (ICU). The framework includes the algorithm for AKI detection from creatinine value and urine output, as well as the prediction of next-day AKI onset from the data collected on the first day in the ICU. The AKI prediction task was implemented for three different granularity levels: predicting an AKI onset of any stage, predicting the AKI onset of stages 2 and 3, and predicting the AKI onset of stage 3, which is the most severe case. Due to the imbalance of the given data, we experimented with several balancing techniques to tackle this problem.

In addition to the classical machine learning approach with manual feature selection, we have also explored an LSTM-based approach applied to the prediction of AKI. Due to the variety of data available for each patient, it is challenging to assess which information could be the best predictor. Thus, the text classification model used unstructured textual data to make predictions.

The extreme gradient boosting (XGBoost) machine learning algorithm, trained on less than 10 thousand patients in imbalanced data settings, achieved better performance than the deep learning text classification model. The latter, in turn, showed the ability to capture meaningful information from the text.

Dedication

With genuine gratitude and warm regard, I dedicate this work to my family and friends. A special feeling of gratitude to my loving mother, who has been showing me love and care by doing everything possible for me to get a good education and have a good life.

I dedicate this thesis to my grandparents, who supported me from an early age and instilled in me a love of knowledge and curiosity.

I also dedicate this work to my close friends, who are always there for me despite the long distance.

Table of Contents

List of Tables viii

List of Figures ix

1 Introduction 2

1.1 Motivation 2

1.2 Research objectives 3

1.3 Overview of the thesis 3

2 Background 5

2.1 Clinical background 5

2.1.1 Defenition of AKI 5

2.1.2 Electronic Health Records 6

2.2 Natural Language Processing in healthcare 6

2.3 Literature review 7

2.3.1 Classical machine learning models 7

2.3.2 Reccurent Neural Networks 12

2.3.3 Transformers 16

3 Methodology 17

3.1 Data 17

3.1.1 MIMIC IV Dataset 17

3.1.2 AKI labels acquisition 17

3.1.3 Cohort analysis 20

3.1.4 Exclusion criteria 22

3.1.5 Predictor variables and input data 24

3.2 Prediction model 25
3.2.1 LSTM structure 25
3.2.2 Architecture of the proposed model 27
3.2.3 XGBoost AKI prediction model 28

4 Experiments and Results 31
4.1 XGBoost models 31
4.1.1 Experimental design 31
4.1.2 Results 32
4.2 LSTM-based model 35
4.2.1 Experimental design 35
4.2.2 Results 36

5 Discussion 40

6 Conclusion 44

References 45

APPENDICES 45

A Python Implementation 51
A.1 Metrics 51
A.2 Libraries 51
A.3 Code 52

List of Tables

2.1 Performances of the models with the best sensitivity and specificity scores grouped by ensemble-based methods. [1] 8

2.2 Performances of ETSM models on ICUC and MIMIC-III [42] 11

2.3 The sensitivity of the model for predicting AKI up to 48 hours with different precision thresholds. 12

3.1 Final cohort statistics description for LSTM-based and XGBoost models. . 21

3.2 Diagnoses and corresponding ICD codes used to compare labels obtained using KDIGO criteria [22] and diagnoses records from MIMIC IV[21]. . . . 23

4.1 Parameters found using grid serach and 5-fold cross-validation for each set of labels. 31

4.2 Results for experiments with XGBoost (XGB) model and different balancing techniques for predicting if a patient will develop AKI during the second day in the ICU. The last column corresponds to the amount of false positive predictions for each true positive. 33

4.3 Additional results of experiments with XGBoost (XGB). The metrics go as follows: precision-recall area under the curve (PR AUC), receiver operating characteristic curve (ROC AUC), and the value for sensitivity when the threshold is chosen for the precision to be equal to 33%. 34

4.4 Results for experiments with different dropout probabilities and embedding size of LSTM-based model for predicting if a patient will develop AKI during the second day in the ICU. The last column corresponds to the amount of false positive predictions for each true positive. 37

4.5 Other metrics describing the results of experiments with LSTM-based model for predicting if a patient will develop AKI during the second day in the ICU. The first two columns describe the dropout probabilities and embedding size of the model. 38

List of Figures

2.1 The architecture of the RNN proposed by Tomasev [40]. 13

3.1 AKI label assigning algorithm based on creatinine part of KDIGO criteria. 18

3.2 An example of AKI label assigning procedure based on urine part of KDIGO criteria for the first day (Day 0) in the ICU for stages 1 and 2. The urine output rate from 12-18h is less than 0.5, so we assign label 1 to stage 1 (according to KDIGO [22]). By the end of the day, the rate is still less than 0.5 for at least 12h, so we assign label 1 to stage 2. Stage 3 label is calculated similarly, except that the time window length is 24h instead of 6h. 19

3.3 The box-plot diagram for patients' weight distribution by AKI stage. . . . 20

3.4 Number of patients having AKI onset on each day during the ICU stay. . . 21

3.5 Distribution of patients having different lengths of ICU stay (a) and different creatinine levels on the first 24h in the ICU (b). 22

3.6 Confusion matrix describing the relationships between labels assigned by the AKI detecting algorithm and diagnoses records from MIMIC IV. *(Notice that samples with AKI of stage 2 or 3 occurred during the first 24h in the ICU are not excluded on this diagram.)* 23

3.7 example of the input data passed to the model. 24

3.8 The structure of a single LSTM cell. 26

3.9 The architecture of the proposed model. 28

3.10 The distribution of the sodium in the training cohort before and after the data normalization using the method described in equations 3.7 and 3.8. . 29

4.1 The quantitative description of the training data fed to XGBoost algorithm, before and after applying sampling techniques. 32

4.2 Top features contributed to the decision of weighted XGBoost model with an undersampled dataset for each subtask. 35

4.3 The experimental design for defining the observation and prediction periods. The medical records from the first 24h in the ICU are used as input for the models. The second 24h are used to define the labels: if a patient developed AKI onset of target stage during this period - it assigned label 1, otherwise 0. 36

4.4 Local interpretations of true negative (a) and true positive (b) examples from the testing dataset for LSTM-based model in experiment 2. 39

List of Abbreviations

AKI	Acute Kidney Injury.
AKIN	Acute Kidney Injury Network.
ATC	Anatomical Therapeutic Chemical.
AUC	Area Under the Curve.
ANN	Artificial Neural Network.
BERT	Bidirectional Encoder Representations from Transformers.
BUN	Blood Urea Nitrogen.
BMI	Body Mass Index.
BPE	Byte-Pair Encoding.
CART	Classification and Regression Tree.
CRF	Conditional Random Field.
EHR	Electronic Health Records.
EMR	Electronic Medical Record.
ETSM	Ensemble Time Series model.
FP	False Positive.
FN	False Negative.
GBT/GBDT	Gradient Boosted Decision Trees.
GFR	Glomerular Filtration Rate.
GRU	Gated Recurrent Unit.
ICU	Intensive Care Units.
IFICF	Indicator Frequency and Inverse Cohort Frequency.
ICD	International Classification of Diseases.
INR	International Normalized Ratio.
KDIGO	The Kidney Disease: Improving Global Outcomes.
LSTM	Long Short-Term Memory.
LIME	Local Interpretable Model-Agnostic Explanations.
MLM	Masked Language Modeling.
NLP	Natural Language Processing.
NB	Naive Bayes.
NSAID	Non-Steroidal Anti-Inflammatory Drugs.
NHS	National Health Service.
NDC	National Drug Code.
PT	Prothrombin Time.
PPT	Partial Thromboplastin Time.

PR AUC	Area Under the Precision-Recall Curve.
ROC AUC	Area Under Receiver Operating Characteristic Curve.
RF	Random Forest.
RNN	Recurrent Neural Net.
SCr	Serum Creatinine.
SVM	Support Vector Machines.
SHAP	Shapley Additive Explanations.
TP	True Positive.
TN	True Negative.
TCN	Temporal Convolutional Network.
WBC	White Blood Count.
XGB	Extreme Gradient Boosting.

Chapter 1

Introduction

1.1 Motivation

Acute kidney injury (AKI) is a medical term for a condition that is characterized as a sharp deterioration in the excretory function of kidneys. Physicians typically diagnose this condition by the increased level of creatinine in a patient's body, which is a final product of nitrogen metabolism. It also can be diagnosed by a sharp decline in urine excretion [5]. AKI is linked to a 4-10-fold increase in the number of deaths which, hence, poses a significant threat to a patient's life. Besides high mortality, this condition increases the length of the hospital stay and associated expenses for patients. Since this injury affects 8% to 16% inpatient admissions, it has a considerable impact on the economy [12].

According to [37], 1% of the National Health Service budget of England is spent on treating AKI patients and other costs related to this syndrome. In the United States, almost half of a million patients suffer from this disorder, and the related treatment costs are from 4.7 to 24.0 billion every year. Moreover, the expenses are increasing with the severity of cases. For example, the difference in cost of the stay for non-AKI patients and high-severe AKI patients who require dialysis is from $\$11,016$ to $\$42,077$ per person [12].

Considerable effort was made to predict and recognize this condition, but only a few predictors are used in clinical practice. This is because their sensitivity and specificity are not high enough, or they detect deterioration too late. Current methods rely on shifts in creatinine (SCr) as a biomarker of renal dysfunction. However, this shift happens after approximately 50% of the glomerular filtration rate (GFR) is already lost, thereby delaying the treatment.

According to the Tomasev Nenad [40], for predictors to be clinically-effective, they should satisfy the following requirements:

- offering meaningful solutions for preventable conditions;
- customizable for individual patients;
- providing enough context information to make informed clinical decisions;

- applicable to a wide range of clinical groups.

The early prediction of kidney decline could be considerable support for clinicians since about 11% of in-hospital deaths could be prevented by promptly recognizing and treating patients at risk. For achieving this goal, early prediction of patient health risks is critical to facilitate the detection of such cases [40].

1.2 Research objectives

Since the task of predicting AKI is quite broad and not straightforward, different studies covered different aspects of it and formulated the objectives differently. For instance, an algorithm can predict an AKI onset without considering its stage (AKI, no AKI) or predict a specific event stage. In terms of the prediction window, an algorithm can make predictions for the next day or the next several months. Besides, an algorithm can be trained on different cohorts, such as patients in ICU wards or undergoing surgery.

Making daily predictions and raising the alert if a patient is at high risk of developing AKI for the next day would assist healthcare professionals. This is because they are limited by time constraints, and it is challenging to cover all the patients in providing a comprehensive assessment of the medical records produced by them.

This thesis aims to explore the following research objectives:

1. using the publicly available Electronic Health Records (EHR) dataset and clinical criteria for diagnosing AKI develop and implement an algorithm for detecting AKI onset and labeling the data;

2. develop a machine learning-based framework for predicting the AKI onset and its stage on the second day of an ICU stay using the data collected on the first day of the stay;

3. explore the different approaches to tackle the imbalance problem in the data within the scope of the task, described in the second objective;

4. explore and evaluate a text classification LSTM-based approach toward the problem described in the second objective by formulating it as a text classification problem and using unstructured EHR data.

1.3 Overview of the thesis

This thesis is structured in the following manner:

- Chapter 1 describes the high overview of the main aspects of the thesis and introduces its research objectives;

- Chapter 2 walks the reader through the clinical definition of AKI, the introduction to Natural Language Processing (NLP) in the healthcare domain, and the EHR concept. Finally, it includes the literature review of related works for predicting AKI using the classical machine learning and deep learning approaches presented in this chapter;

- Chapter 3 describes the methodology addressing the research objectives. First, it focuses on the data we work with: analyzes the cohort, describes how we excluded samples from the final cohort, explains what prediction variables were used, and data preprocessing. Second, we describe and analyze the label acquisition process. Finally, the prediction model architecture is presented for predicting AKI as an text classification task, as well as the classical machine learning approach for AKI predicting task;

- Chapter 4 describes the experimental design for the conducted experiments and presents the results of these experiments;

- Chapter 5 contains a discussion of the results and findings from the experiments; the limitations and contributions of this work; and suggests the directions for future research;

- Chapter 6 summarizes the work.

Chapter 2

Background

2.1 Clinical background

2.1.1 Defenition of AKI

There are several criteria to determine AKI, which use urine output or serum creatinine level. Particularly the Acute Kidney Injury Network (AKIN) criteria [33], and the RIFLE (risk, injury, failure, loss, end-stage) criteria [6]. Later, the new definition was invented by merging the RIFLE and AKIN criteria: KDIGO (the Kidney Disease: Improving Global Outcomes) criteria [22].

According to KDIGO, an AKI case occurs when a patient meets at least one of the conditions listed below:

- Within 48 hours, SCr level has increased by 0.3 mg/dL or greater;
- Within the last seven days, SCr level, has increased to 1.5 times baseline or greater;
- Within 6 hours, urine output is less than 0.5 mL/kg/h.

In addition, KDIGO provides definitions for three stages of the severity of AKI:

Stage	Creatinine	Urine output
1	$1.5 - 1.9$ times more than baseline or $\geq 0.3mg/dL$ increase	< 0.5 mL/kg/h within 6 h
2	$2 - 2.9$ times more than baseline	< 0.5 mL/kg/h within 12 h
3	3 times more than baseline or $\geq 4mg/dL$ or renal replacement treatment is started	< 0.3 mL/kg/h within 24 h or anuria for 12 h or more

Several conditions often pose AKI, for example, having a slow blood flow or direct damage to the kidneys [1]. A recent study conducted in 2020 showed that 19% of adult

[1] https://www.elsevier.com/books/ferris-clinical-advisor-2020/ferri/978-0-323-67254-2

people hospitalized due to active SARS-CoV-2 (severe acute respiratory syndrome which provokes COVID-19 disease) experienced AKI after or during the disease [15]. The rate of death among AKI patients was 47%. Moreover, the study showed that neutrophil count at admission and other factors such as age, chronic kidney disease, and mechanical ventilation are also risk factors for AKI. The other study from Mexico showed that morbid obesity of patients with COVID-19 was linked to a higher risk of AKI and death [14].

2.1.2 Electronic Health Records

The term electronic health records (EHR), also known as personal health records (PHRs) or electronic medical records (EMRs), is extensively used in the clinical community. It describes a conception of extensive cross-institutional long-term information collection about patients' health. EHR is a database comprised of various types of data related to a patient's physical condition, treatment, and health in general [19]. Particularly demographic information such as age and gender, diagnoses, surgeries, and laboratory tests. The patient actively participates in their treatment by accessing, adding, and updating it, thereby supporting care.

EHR consists of results and a summary of interactions between healthcare providers and patients. Therefore, these databases mirrored the style of healthcare professionals, their knowledge, and the specifications of the systems they work with. To systematize and unify the structure of EHR the Institute of Medicine in 1991 [20] determined the golden standard and described functions of EHR.

Considering the comprehensiveness and informativeness of EHR data, Deep Learning predictive models are expected to boost personalized medicine and increase the quality of healthcare.

2.2 Natural Language Processing in healthcare

NLP is a field where Artificial Intelligence, Computer Science, and Linguistics meet to allow computers to learn how to understand human languages. Classical supervised machine learning techniques such as Support Vector Machines (SVM), Conditional Random Field (CRF), and Maximum Entropy are used for various NLP tasks[10][25][35]. Namely, breaking a text into pieces (tokenization) and named entity recognition which is identifying entities (e.g., people's names, cities, mobile numbers), to name a few. Another example is text classification - analyzing a piece of text and assigning it to a specific category, for example, fake or real (fake news detection), positive or negative or neutral (sentiment analysis), or categorizing documents by topic.

Although these machine learning models are still widely used for some tasks, deep neural nets currently have the largest capacity for working with high-dimensional text data. Thus, Recurrent Neural Nets (RNNs) such as Gated Recurrent Unit (GRU) or Long short-term memory (LSTM) models give a sense of historical information depending on the position of a word in a sentence. Another example is transformers, which utilize the attention mechanism to identify the most meaningful information in the text. The latter is currently showing state-of-the-art results in the NLP field[41].

Recent studies showed that NLP techniques could be applied to medical data due to their similarity to textual data. For example, suppose a patient has several visits to a hospital. Each diagnosis made on a particular visit can be seen as a word, the sequence of the diagnoses made during the visit as a sentence, and all history of diagnoses of this patient as a document [27]. Similarly to diagnoses, other information related to a patient's visit (e.g., prescribed medications, vital signs, demographics) might also be used. Deep neural models can be trained using this information to predict a patient's future health trajectories and risk of a specific disease. This can help obtain meaningful insights and support clinicians in decision-making. Above mentioned data can be acquired with the help of EHR systems, which are currently adopted by roughly 95% of critical access hospitals [2].

2.3 Literature review

2.3.1 Classical machine learning models

Bell et al [3] used only four variables for the binary logistic regression model to predict the AKI stage. Namely, they used age, estimated glomerular filtration rate (eGFR) category, previous diabetes mellitus, and heart failure diagnoses. They acquired the weighted scores identifying the stage of AKI by measuring the change in SCr. They defined baseline SCr as a median value from 8 to 365 days before the given SCr measurement. If no measurements are available for this period, they defined the baseline as the lowest level of SCr from 0 to 7 days before the given measurement. In case of no SCr baseline measurement available, the shift of SCr more than $26\mu mol/L$ in 48 hours was taken as AKI stage 1.

The C-statistics of the model developed on 273,450 patients from the Tayside region of Scotland was 0.80 (0.80-0.81) with ($95\% CI$). Moreover, it was evaluated on two more cohorts: from Kent, UK, and from Alberta, Canada. The performance was 0.76 (0.75-0.76) in the large Canadian cohort with 1,173,607 patients, and 0.71 (0.70-0.72) in the Kent cohort with 219,091 patients. This quite simple score with only four variables is one of the few externally validated predictors with such a considerably big group of patients. Since the model did not use any medications, it is a reasonable future research goal that could improve the model's performance.

Many existing methods use logistic regression for predicting AKI cases. However, there is a limitation in using a logistic regression approach. Logistic regression assumes a linear dependence of all predictor and dependent variables, weakening the model's discrimination power, i.e., the ability to discriminate between positive and negative samples.

Sheikh S. Abdullah et al. [1] studied the risk of a patient being readmitted to a hospital with AKI during the next three months after the discharge. They conducted experiments on a cohort of EHR of $905,442$ older patients (from 65 years and older) where

[2] `https://www.definitivehc.com/blog/hospital-ehr-adoption`

Ensemble-based methods	**Machine Learning Techniques**	**Sensitivity**	**Specificity**	**ROC AUC**
XGBoost	Linear boosting	0.86	0.77	0.84 ± 0.033
	Tree boosting	0.89	0.81	0.88 ± 0.031
RUSBoost	SVM (sigmoid)	0.90	0.79	0.88 ± 0.029
	SVM (radial)	0.71	0.87	0.85 ± 0.034
Undebagging	SVM (sigmoid)	0.89	0.71	0.85 ± 0.034
	SVM (radial)	0.79	0.90	0.86 ± 0.033
SMOTE-Bagging	SVM (radial)	0.90	0.74	0.86 ± 0.029

Table 2.1: Performances of the models with the best sensitivity and specificity scores grouped by ensemble-based methods. [1]

they evaluated the performance for 31 AKI prediction models. Each model was a combination of two or more out of the following classifiers - SVM with four kernels (linear, radial, sigmoid, polynomial), logistic regression, naive Bayes (NB), classification and regression tree (CART), C5.0. In addition, they used two sampling methods - SMOTE and undersampling, and three ensemble methods, such as XGBoost, Bagging, and Boosting.

Since data was highly imbalanced (with only 5993 AKI patients), the traditional machine learning techniques failed to map the positive class correctly. This is because they aim to reduce the total error, and the positive class does not contribute much to it. Therefore, four sampling and ensemble methods and their combinations were used to address this issue. All of the models were assessed using a 10-fold cross-validation method.

The best performance in terms of ROC AUC amounted to 0.88. This result was achieved using (1) a combination of SVM with sigmoid kernel and RUSBoost and (2) XGBoost with tree boosting. Considering that it is a patients outcomes prediction task, a sensitivity (the ratio of actual positive samples that are predicted correctly as such (*i.e.*, $SN = TP/(TP+FN)$) is more important than a specificity (the ratio of actual negative samples that are predicted correctly as such (*i.e.*, $SP = TN/(TN+FP)$). SVM with sigmoid and radial kernels combined with RUSBoost and SMOTE-Bagging achieved a sensitivity of 0.90, which was the best result among all models. Table 2.1 shows the performances of the models used in the study.

The work of Sheikh S. Abdullah has some limitations that should be addressed in future research. *First*, the training and testing were conducted on the cohort of old patients from Ontario. Therefore, models will not generalize well in the case of different locations and ages of a patient. *Second*, records for patients who did not have complete demographic information were excluded from the final cohort, which may affect the performance due to the high probability of missing the rare cases. *Third*, the work does not imply any ranking system for predictors and only establishes the most significant ones because of the different approaches to identifying feature importance. *Fourth*, the models used diagnosis codes to define AKI cases. Hence, it is not capable of identifying the severity of a case.

Ke Lin et al.[28] conducted a study where they showed that the Random Forest model outperforms three other methods in predicting the probability of a patient's in-hospital death given that they have AKI. The final cohort consisted of 19,044 AKI patients with $2,586$ patients who died in the hospital. They used several methods to estimate the mortality of AKI patients: Random Forest (RF) with 1000 trees, Artificial Neural Network (ANN) with one hidden layer of 10 nodes, and SVM.

During the data preprocessing step, they filled missing values with their corresponding means and excluded the outliers. For the time-stamped measurements, the maximum and the minimum values during the first 24 hours of a patient's stay in the ICU were used as two parallel inputs. The average performance measures were generated after five training rounds with 5-fold cross-validation.

The RF model was compared to three other models and outperformed them in terms of ROC AUC ($0.866, 95\%CI : 0.862 - 0.870$) and Brier score ($0.085, 95\%CI : 0.084 - 0.086$). Both RF and SVM models had the best discrimination power. It is worth noting that the RF model marginally overstated the death rate of low-risk patients while underestimating the death rate of high-risk patients. Both models slightly underestimated the death rate of patients.

The given study has several strengths: the availability of the data, the credible performance of the models, and the use of 5-fold cross-validation to verify the results. In addition, the variables used for prediction are clinically available; therefore, the model can be used in clinical practice. Moreover, the RF algorithm allows tracking each variable's importance for the model. Consequently, this approach can achieve good interpretability.

However, this study has a few drawbacks. *First*, the data comes from one distribution: patient data collected in one medical center from 2001 to 2012. The model might not perform well when working with natural clinical settings or other datasets. *Second*, some biomarkers related to AKI are not present in the dataset and were not included in the predicting variables [29].

Cheng et al[9] conducted another revealing study in 2017 in China. They showed that the RF model had better performance in the AKI prediction than the other machine learning methods, e.g., Logistic Regression and AdaboostM1. The study aimed to answer three questions:

1. How much data prior to admission improves the prediction performance?
2. What is the earliest time of predicting AKI with good predicting power?
3. How much do the specific predictors contribute to the model performance?

The final cohort consisted of $48,955$ admissions ($33,703$ patients) of patients aged 18 to 64 admitted to an academic hospital (the University of Kansas Medical Center – KUMC) and stayed there for at least two days from November 2007 to March 2016. They excluded patients with less than two measurements of SCr and those with eGFR lower than $60mL/min/1.73m2$ or SCr $1.3mg/dL$ during the first 24 hours of hospital stay. The final cohort included $48,955$ admissions ($33,703$ patients) with $9\%(4,405)$ AKI admissions. The last laboratory test values and vitals before a prediction point were divided into three categories: present and normal, present and abnormal, and unknown. Vital signs and missing laboratory tests were categorized as 'unknown' because deciding not to perform specific measurements can be potentially informative.

For the first objective, they were iteratively changing the lower bound of the prediction window to see how much the amount of the prior data affects the performance of the model. The best AUC values of prediction models on data collected before and after hospital admission were achieved by RF for predicting the AKI case 1 day before the event, which showed that data prior to admission does not improve performance.

To address the second objective, they changed the upper bound of the prediction window by moving the prediction point further from the event (AKI or discharge from the hospital). They were aiming to learn how well an AKI event can be predicted and how early. The results showed that model performance drops when the time window between prediction time and AKI event time expands. The Random Forest's AUC dropped from 0.76 at 1-day prior to 0.67 at 5-days prior. However, the AdaBoostM1 had a slightly better AUC, with the time-to-event horizon extended to four or five days before the event. In addition, the Logistic Regression had slightly better precision and recall for one and two days prior settings, respectively.

For the third objective, they combined the predictors in several groups, then iteratively removed the groups and trained the model with the rest groups' predictors. This experiment showed that the medications and comorbidities had the most impact on the 1-day prior AKI prediction performance and the demographics variables affected the performance the least.

According to recommendations from the 15th Acute Dialysis Quality Initiative (ADQI) consensus conference in 2016, the prediction can be clinically valuable if it was made from 48 to 72 hours before the occurrence of AKI event [38]. The study showed that the RF model could predict the AKI 2-days and 3-days before and achieve AUC of 0.73 and 0.70, respectively. However, this study has several limitations. *First*, the final cohort consisted of relatively young patients. Taking into account that old patients are more at risk of AKI due to longer exposure to chronic diseases and nephrotoxins, the model might not generalize well. *Second*, the only patients with normal SCr level and with eGFR of at least $60mL/min/1.73m$ were incorporated in the final cohort. *Third*, the study did not consider the time-dependence of the comorbidity variables. Therefore, the comorbidities that developed during the admission might be misrepresent. *Finally*, the urine output measurements were not included in the predictors, although it is one of the defining criteria of AKI.

Wang et al.[42] proposed the Ensemble Time Series model (ETSM) based on XGBoost for predicting AKI onset 24 and 48 hours in advance. Vital signs, laboratory test results, and medications were used as predictors. Combinations of drugs taken by patients on each day of admission during the observation period were grouped into sequences. In turn, the sequences were used to calculate Indicator Frequency and Inverse Cohort Frequency (IFICF). They showed the following top 10 important combinations for predicting AKI:

- norvancomycin

Dataset	**AUC**	**Sensitivity**	**F1-score**	**AP**
24h prediction				
ICUC	0.81	0.75	0.58	0.59
MIMIC-III	0.95	0.95	0.96	0.98
48h prediction				
ICUC	0.78	0.68	0.44	0.41
MIMIC-III	0.95	0.98	0.98	0.98

Table 2.2: Performances of ETSM models on ICUC and MIMIC-III [42]

- ciprofloxacin lactate and sodium chloride injection
- indometacin enteric-coated tablets
- piperacillin sodium/tazobactam sodium
- ibuprofen
- ceftazidime for injection
- cefathiamidine for injection
- aztreonam for injection
- naproxen

The impact of medication information on the model performance was shown by training the same model without adding drug records. The difference in AUC between the two models amounted to 0.07 when predicting AKI 24 hours in advance (AUC 0.74 and AUC 0.81).

This study used two datasets to build and evaluate the models: ICUC (ICU in China), which is quite imbalanced with a prevalence of negative samples, and a more balanced MIMIC-III. They used 11,501 and 10,921 samples to predict AKI onset in 24 and 48 hours from ICUC, as well as 46,593 and 30,217 samples from MIMIC-III, respectively. The random undersampling technique yielded better results than random oversampling and cost-sensitive XGBoost. Table 2.2 shows the model's performance on both datasets for 24 and 48 prediction windows.

The limitation of this study is that the model was trained on ICU patients' data only. Therefore generalizability of the model on patients from other departments is subject to further investigation.

Labels	25% Precision	33% Precision	40% Precision	75% Precision
any AKI	0.68	0.56	0.47	0.12
stages 2 or 3	0.78	0.71	0.65	0.28
stage 3	0.89	0.84	0.8	0.4

Table 2.3: The sensitivity of the model for predicting AKI up to 48 hours with different precision thresholds.

2.3.2 Reccurent Neural Networks

Tomasev et al.[40] trained the RNN on a dataset of 703,782 adult patients from the US Department of Veterans Affairs (VA). The ratio of KDIGO AKI admissions in the dataset was 13.4%. The model provided estimates for the risk of AKI during the following 48 hours at each time step and the corresponding uncertainty degree. The model achieved 92.1% of ROC AUC and 29.7% of PR AUC, corresponding to 55.8% inpatient cases of any severity of AKI that were predicted correctly within a given time window. Regarding the stage 3 AKI, 84.1% of the cases were correctly predicted up to 48 hours. Table 2.3 shows the sensitivity of the model for predicting different stages of AKI with fixed precision thresholds.

During data preprocessing, each patient was represented by a sequence of blocks. Each block consisted of diagnoses, procedures, prescriptions, medications, laboratory results, health factors, vital signs, and note titles recorded within the 6 hours window. In other words, each day consisted of four six-hour blocks and one additional block of events with no time recorded. The model additionally might take an embedding of the preceding 48 hours and a 6-month or 5-year history.

All variables were initialized with normalized initialization (Xavier) and then trained using the Adam optimizer. The model's architecture is shown in Figure 2.1. The training process was implemented as follows:

1. the embedding layers of the model transform the input;
2. the embeddings are passed into an RNN with stacked multiple-layer where all layers are connected;
3. the output of the RNN is passed into a prediction module which provides the AKI probability over eight feature prediction windows.

The given study has several strengths. *First*, the features used for predictions were chosen based on the opinion of six experienced clinicians. *Second*, the uncertainty for predictions was provided by extensive training of an ensemble consisting of 100 models, a set of hyperparameters that are initialized with different seeds and fixed. *Third*, validation set performance was used to decide in favor of the proposed model.

However, this study also has several limitations. *First*, patients with less than one year of available EHR data before admission were excluded. Hence, it is unclear how the

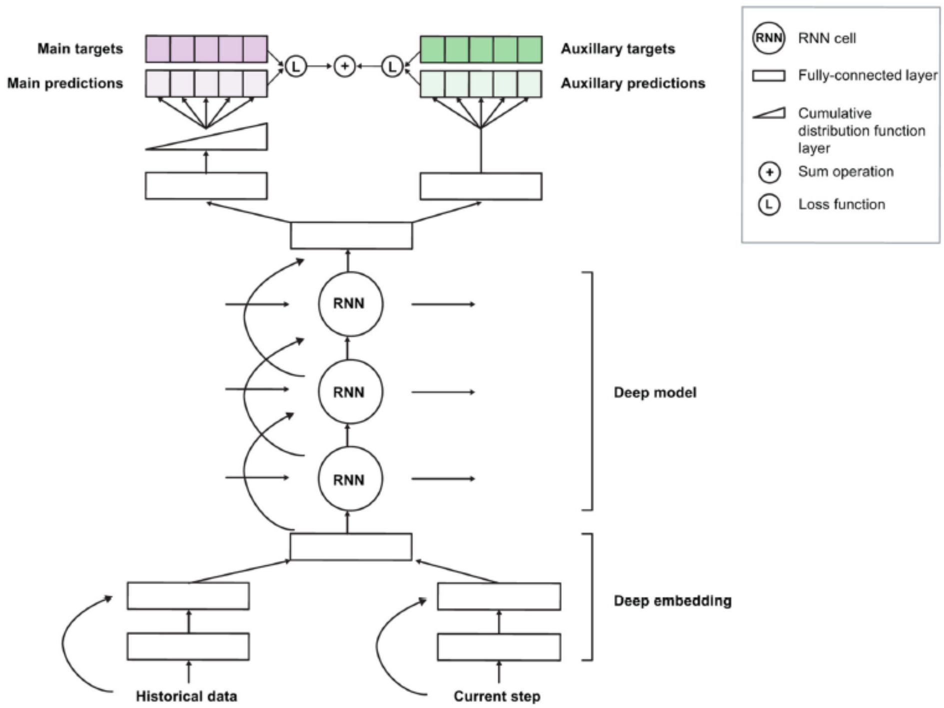


Figure 2.1: The architecture of the RNN proposed by Tomasev [40].

model will perform on a patient without prior medical history. *Second*, the AKI cases were defined based only on the creatinine level, while the urine criteria were not considered for identifying AKI. Consequently, it can lead to an increasing number of false negative diagnoses.

Nina Rank et al. [34] compared the performance of experienced clinicians and the RNN model trained on a balanced set of 2224 admissions. This model surpassed clinicians by a substantial margin ($AUC = 0.901 vs. 0.745, p < 0.001$) and provided real-time predictions (every 15 minutes) of AKI during the first seven days after cardiothoracic surgery. Compared with clinicians, the RNN had a higher sensitivity, which was 0.851. Moreover, sensitivity reached a maximum of 0.971 on the interval 2-6 hours before the onset. The lower bound of achieved sensitivity was 0.750 on the 48–168 hours interval before the onset. The Adam optimizer and learning rate of 0.001 were used for training.

It is worth noting that this study correctly defined 30% of all AKI cases using the urine criteria in addition to creatinine criteria. It makes the given study different, for example, from the work of Cheng et al.[9] and Tomasev et al.[40], where they used only creatinine criteria. Nina Rank et al. showed that 11% of the training population and 12% of the testing population in their study satisfied the urine output criteria and would have been labeled as false negative without using it. In contrast to Tomasev's study, the only information generated after the admission was used, which makes it useful in a natural clinical setting where patients may not have a previous medical history.

Kipyo Kim et al [23] conducted another study on predicting AKI in 2021. They proposed a prediction model based on a stacked RNN structure using the information collected between 2013 and 2017 in two hospitals in Korea. For training, they utilized the data of $69,081$ patients from the first hospital and the data of 7675 patients from the same hospital for internal validation. In addition, they externally validated the model on the other cohort of $72,352$ patients from the second hospital. The final cohort comprises patients older than 18 with at least 48 hours of recorded information. The patients falling under the following criteria were excluded from the final cohort:

- no baseline SCr measurements;
- SCr greater than 4.0 mg/dL or eGFR less than 15 mL/min/1.73 m2;
- The final stage of kidney disease at the time of the admission;
- absence of laboratory measurements used in the model;
- absence of Body Mass Index (BMI);
- absence vital signs measurements;
- an AKI diagnosis on the first day of admission.

The model used 107 predicting variables divided into static and dynamic categories. Comorbidities and demographics were defined as static variables, whereas vital signs, laboratory tests, and clinical conditions were defined as dynamic. Categorical features were coded as numbers and percentages. For continuous variables, they used min-max normalization. It is worth noting that the training set had a higher number of patients and a higher average age than the external validation set. The importance of the variables was estimated using the Shapley Additive Explanations (SHAP) algorithm [30], as well as other model-agnostic methods. The importance of several dynamic variables, such as white blood cell count, diuretic use, and pulse rate, increased with time. Moreover, SCr levels, white blood cell counts, and vital signs had the highest impact the day before the prediction point. The hemoglobin and albumin, in contrast, had the highest impact 5-6 days before the prediction.

The AKI cases were defined according to KDIGO creatinine criteria, where the baseline SCr corresponded to the minimum value within two weeks before the admission or the minimum value measured between 90 or 180 days prior to the admission, or the value measured on the first day of the admission. The overall architecture consists of two models: model 1 predicts the AKI onset within seven days of the selected time point, and model 2 predicts the SCr values up to 48 hours from the present. Model 1 has many-to-one architecture, where the input is seven sequences (corresponding to the sliding window length), and the output is whether AKI occurs in the next seven days from the given time point. Model 2, in turn, has many-to-many architecture, where the input is seven sequences, and the output is the three predicted trajectories of SCr values after 1, 2, and 3 days from the present.

Model 1 achieved the AUC of 0.84 (external validation) and 0.88 (internal validation) for any AKI development stage, along with AUC of 0.90 (external validation) and 0.93 (internal validation) for stage 2 and 3 severity of AKI. Model 2, in turn, obtained mean-squared errors of $0.04 - 0.09$ and $0.03 - 0.08$ for patients at higher risk and lower risk of AKI, respectively. The limitation of this study is that the urine criteria were not used for identifying AKI. Another limitation is that the important in medical domain metrics such as sensitivity and precision were not reported. One of the main strengths of the given study is that the model was externally validated on an extensive cohort of patients from the other hospital, which makes the results more trustworthy.

Chen et al. [8] made use of records from ICU stay from the MIMIC-III dataset and attention-based neural network to predict AKI onset, AKI stage, and AKI onset time interval in three scenarios, namely:

1. *Case 1*, when they use the first 24 hours data from ICU stay to predict the next 24 hours outcome in patients using onset intervals set to 12 hours;
2. *Case 2*, when they use the first 24 hours data from ICU stay to predict the next six days outcome in patients using onset intervals set to 24 hours;
3. *Case 3*, when they use data from the first 48 hours of ICU stay to predict the next five days' outcome in patients using onset intervals set to 24 hours.

The model consists of an LSTM encoder, followed by an attention module and a fully-connected decoder. Hidden states outputted by the encoder are passed to the attention function to produce a context vector which, in turn, is used by the decoder model to generate predictions. In addition, a Temporal Convolutional Network (TCN) is used to predict future values of the temporal features such as the laboratory test results and the vital signs.

The experiments showed that the proposed attention-based deep learning model outperformed logistic regression, random forest, and the Gradient Boosted Trees (GBTs), as well as LSTM model in the AKI case prediction task. They also showed that the model with predicted future values had a 0.04 boost in AUC compared to the one without it. Moreover, the model reached AUC 0.85 in experimental case 3. In contrast, it reached AUC 0.82 in experimental case 2, which suggests that increasing the observation window from 24h to 48h positively affected the predictive power of the model. When predicting onset intervals in experimental cases 2 and 3, the model had a better AUC (around 0.7) for predicting the first two 24h onset intervals, compared to predicting the rest onset intervals (AUC 0.52).

The given study also evaluated the importance of the different features for predicting AKI onset using the weights from the attention function. The following variables were found to be the most important for the model to make predictions: urine output, creatinine, blood urea nitrogen (BUN), blood glucose, sodium, as well as non-steroidal anti-inflammatory drugs (NSAIDs), and lipid lowering medications. Similarly, the following comorbidities were also highly-correlated with AKI: diabetes, peripheral vascular, hypertension, and congestive heart failure.

2.3.3 Transformers

In their work, Yikuan and Shishir proposed a BERT-based transformer model named BEHRT, which predicts the future diagnoses for subsequent hospital visits from the medical history [27][13]. The authors built and evaluated the model on more than 1.6 million patients from Clinical Practice Research Datalink (CPRD) dataset from the UK. The BEHRT model takes as an input the text sequence of ICD-codes (i.e., words) corresponding to patients' diagnoses from previous hospital visits.

The model was pre-trained using the masked language modeling (MLM) technique when some of the words in a sentence are hidden from the model, and the objective is to predict these words. In the fine-tuning stage, the authors defined three downstream tasks to train and evaluate the model performance: prediction of the diagnoses in the next visit, during the next six months, and the next 12 months.

BEHRT outperformed previous state-of-the-art by 8.0–13.2% in average precision scores among all three downstream tasks. In addition, the model showed the ability to find relationships between different diagnoses due to its self-attention mechanism.

One of the crucial contributions of this study is experimenting with the novel way of approaching a healthcare problem as an text classification task using the similarity of EHR and natural language. This is different from the traditional way of utilizing EHR in disease prediction models, where manually selected concepts are represented as input features [2]. This approach opens up opportunities to utilize the most available information without relying on manual feature selection. However, one limitation of this study is using only diagnoses without considering other concepts related to patients' health, such as prescribed medications, laboratory test results, and procedures.

Chapter 3

Methodology

3.1 Data

3.1.1 MIMIC IV Dataset

The Medical Information Mart for Intensive Care (MIMIC-IV) v2.0 dataset [21] was used for experiments. This dataset contains EHR collected by the clinical information system for critical care from patients admitted to the critical care units of the Beth Israel Deaconess Medical Center between 2008-2019. The database includes information regarding patients' hospital encounters, such as admissions, demographics, transfers, diagnoses, laboratory test results, vital signs, prescribed medications, and procedures.

3.1.2 AKI labels acquisition

The clinically effective predicting algorithm should identify if a patient is at a high risk of AKI in place, i.e., during the current visit to the hospital. Therefore, the model must make fine-grained predictions timewise. Since the diagnoses are inserted into the MIMIC-IV database at the end of a patient's visit or during the discharge, they do not have corresponding timestamps. Due to that, straightforward use of the AKI diagnoses as labels for each day in the ICU is impossible.

Thus, the day-level labels were obtained directly from laboratory test results of creatinine measurements and values of urine output volume using KDIGO criteria [22]. The label is represnted by 0 or 1 for a given prediction time window (24 hours) for a given sample (an ICU stay). The labels based on creatinine criteria were obtained using the following procedure:

1. For every day, two baseline values of creatinine corresponding to two AKI criteria were calculated (the baseline 1 - the *minimum* value within the last two days and the baseline 2 - the *minimum* value within the last seven days). If there is no SCr measurement for the last seven days, a median creatinine for the last 360 days was

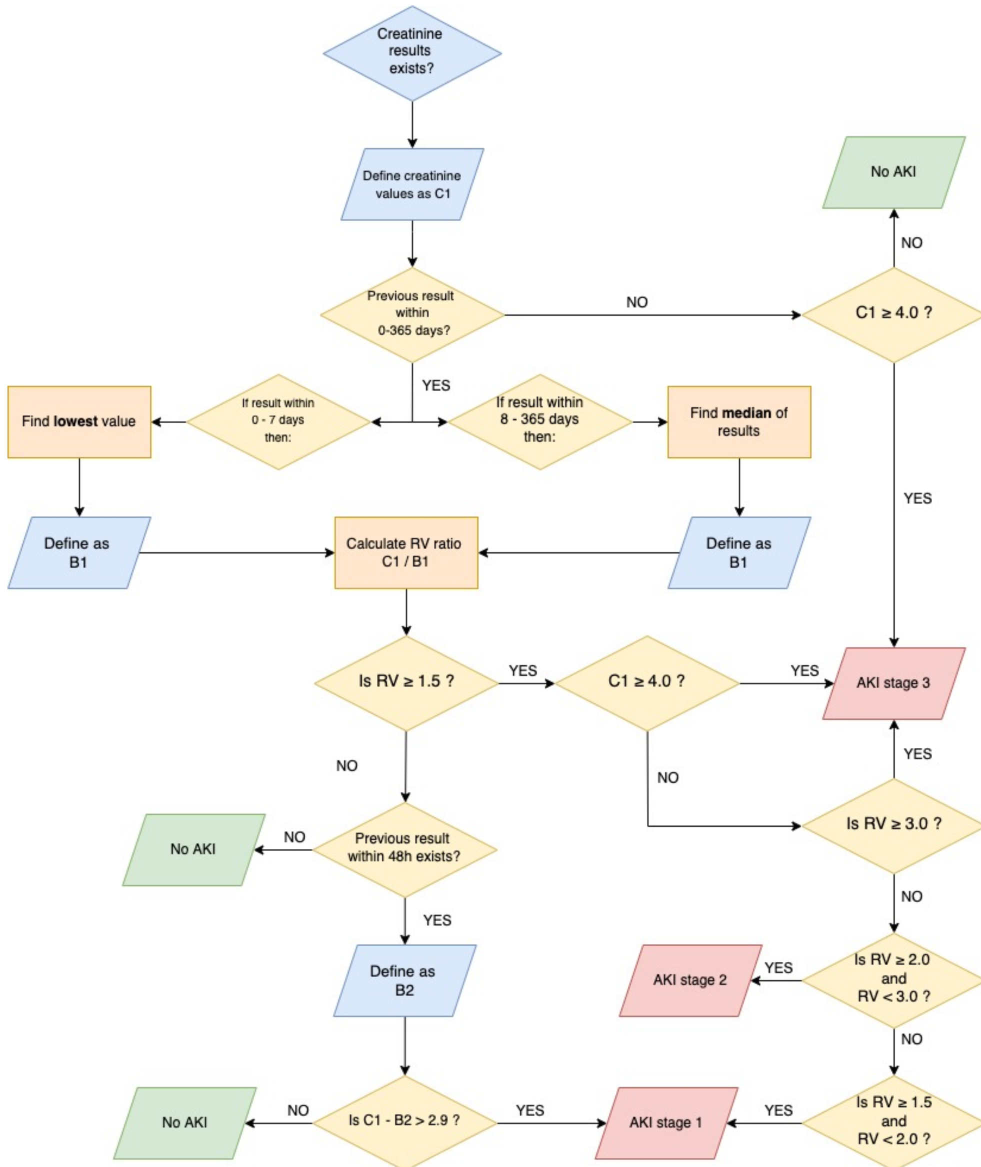


Figure 3.1: AKI label assigning algorithm based on creatinine part of KDIGO criteria.

used as baseline 2. Otherwise, if there are no baseline values were not available (e.g., a patient does not have any creatinine measurements within the last 360 days), the NaN value has been assigned to this baseline.

2. Then, if a patient has creatinine measurement on a given day (C1) more or equals 4.0 mg/dL, we assign label 1 to stages 1,2 and 3 (since satisfying AKI of stage 3 criteria implies satisfying AKI of stages 1 and 2 criterias).

3. Next, if baseline 1 is available, we calculate the ratio between C1 and baseline 1 (RV), and if RV is more or equal to 3.0, we assign label 1 to stages 1,2, and 3. If RV is less than 3.0 but more or equal to 2.0, we assign label 1 to stages 1,2, and 0 to stage 3. Finally, if RV is less than 2.0 but more or equal to 1.5, we assign labels 1 to stage 1

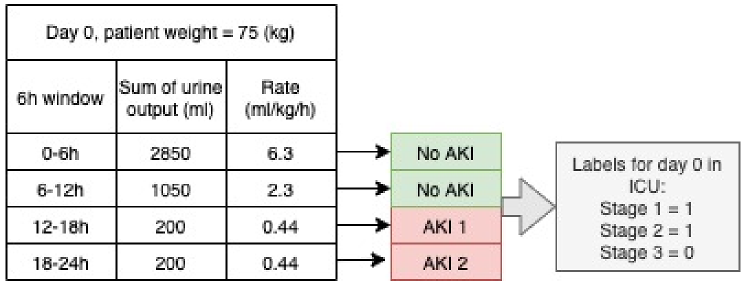


Figure 3.2: An example of AKI label assigning procedure based on urine part of KDIGO criteria for the first day (Day 0) in the ICU for stages 1 and 2. The urine output rate from 12-18h is less than 0.5, so we assign label 1 to stage 1 (according to KDIGO [22]). By the end of the day, the rate is still less than 0.5 for at least 12h, so we assign label 1 to stage 2. Stage 3 label is calculated similarly, except that the time window length is 24h instead of 6h.

and 0 to stages 2 and 3. Otherwise, we assign 0 to all stages (for now).

4. If baseline 2 is available and the difference between it and C1 is more than 0.29, we assign label 1 to stage 1.

5. If no baseline values are available, we assign 0 to all stages.

Figure 3.1 shows the diagram describing the algorithm for obtaining creatinine-based labels. This algorithm is a modified version of the one that originated from The National Health Service (NHC) of England [1].

Similarly, the urine output values were analyzed to calculate labels for AKI stages. The values of urine output were aggregated to get the sum in milliliters of outputs for each of 6h time windows, 12h time windows, and 24h time windows. Then, these values were divided by the patient's weight in kilograms and the number of hours in a specified time window. Then, the obtained values of urine output rates were analyzed to assign the labels:

1. if the rate is less than 0.5 ml/kg/h for six hours, stage 1 was assigned label 1, otherwise label 0;

2. if the rate is less than 0.5 ml/kg/h for 12 hours, stage 2 was assigned label 1, otherwise label 0;

3. if the rate is less than 0.3 ml/kg/h for 24h, stage 3 was assigned label 1;

4. if there are no urine output values recorded for a given day, assign label 0 to all stages.

An example of the above-mentioned process is shown in Figure 3.2.

The final labels for each day in the ICU unit were calculated by combining both: labels obtained using creatinine criteria and urine output criteria. If any criteria were satisfied for any stage, we would consider it as the AKI onset of the given stage.

[1] `https://www.england.nhs.uk/akiprogramme/aki-algorithm/`

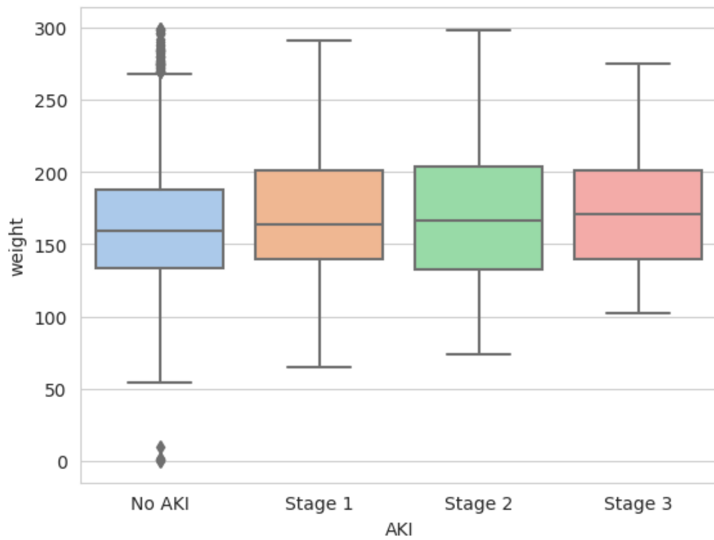


Figure 3.3: The box-plot diagram for patients' weight distribution by AKI stage.

Subtasks formulation. Finally, three sets of labels were constructed: 1) any patient satisfying stage 1 of AKI according to KDIGO [22] during the second day in the ICU was assigned label 1, otherwise, they were assigned label 0; 2) any patient satisfying stage 2 of AKI during the second day in the ICU was assigned label 1, otherwise label 0; 3) any patient satisfying stage 3 of AKI during the second day in the ICU were assigned label 1, otherwise label 0. It is worth noting that the higher stage's satisfaction implies the lower stages. Therefore, the first set of labels could also be defined as *'any stage'* of AKI, the second set of labels could be seen as *stages 2 and 3* of AKI, while the third set of labels represent the patients having *the stage 3 of AKI*. We consider all three sets of labels as different subtasks of the AKI prediction task. Each has different granularity and was evaluated separately later in the results section.

3.1.3 Cohort analysis

The MIMIC-IV version 2.0 dataset contains information on 53,569 patients with 76,943 ICU stays. Most of the patients stay in ICU for one to four days. Figure 3.5a shows the distribution of patients' length of stay. Daily AKI status calculated via the algorithm mentioned above is available for 33,806 ICU stays. Figure 3.4 shows the number of patients having AKI at each stage during the ICU stay. Most of the AKI onsets occurred during the ICU's first three days. The peak is in the first 24h, indicating that kidney failure was the reason for these patients' transfer to the ICU. From the creatinine level during the first 24h in the ICU (Figure 3.5b), we can notice many patients having creatinine higher than 1.2 mg/dL, while the normal creatinine level for adult women is 0.59 to 1.04 mg/dL, and for adult men is 0.74 to 1.35 mg/dL [2]. The most common stage of AKI in the cohort is 1.

In the final cohort (after filtering out samples according to all exclusion criteria), the

[2] `https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646`

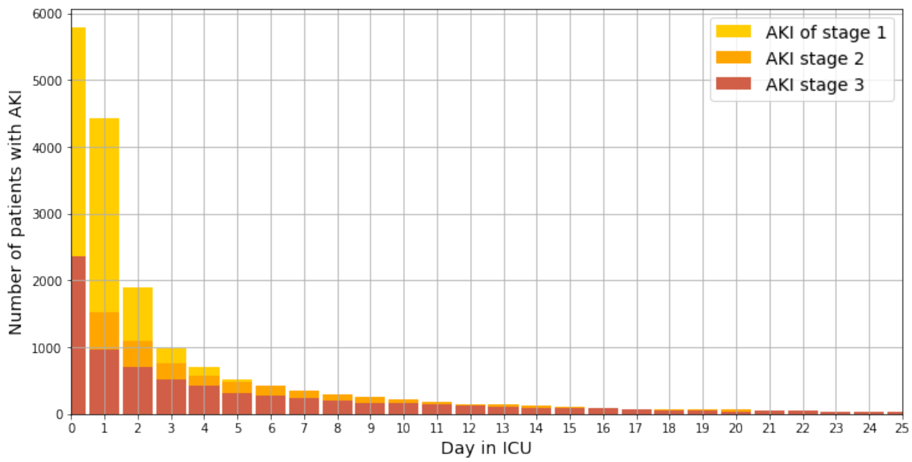


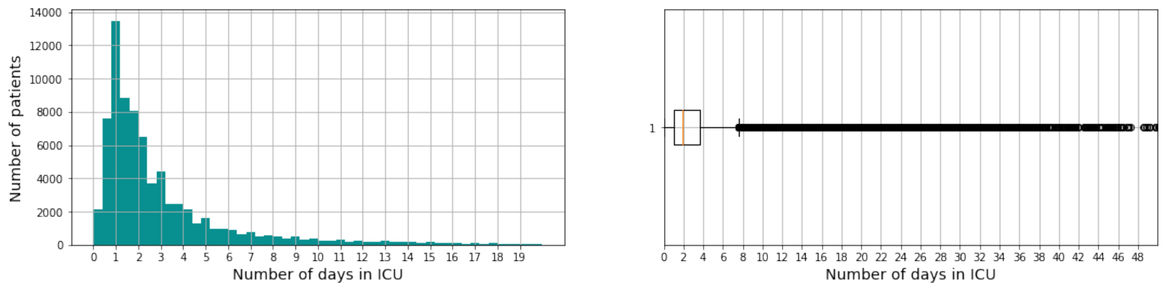
Figure 3.4: Number of patients having AKI onset on each day during the ICU stay.

Statistic	**LSTM-based**	**XGB**
n of samples - total	14,282	13,128
n of samples - train	11,425	10,487
n of samples - test	1,428	1,328
n of samples - val	1,429	1,313
Stage 1 AKI	0.09	0.11
Stage 2 AKI	0.03	0.04
Stage 3 AKI	0.01	0.01
Women	0.46	0.45
White race	0.63	0.64
Black/African american	0.14	0.06
Other ethnicity	0.23	0.3
Age - Q2 (median)	52	65
Age - Q3	70	77

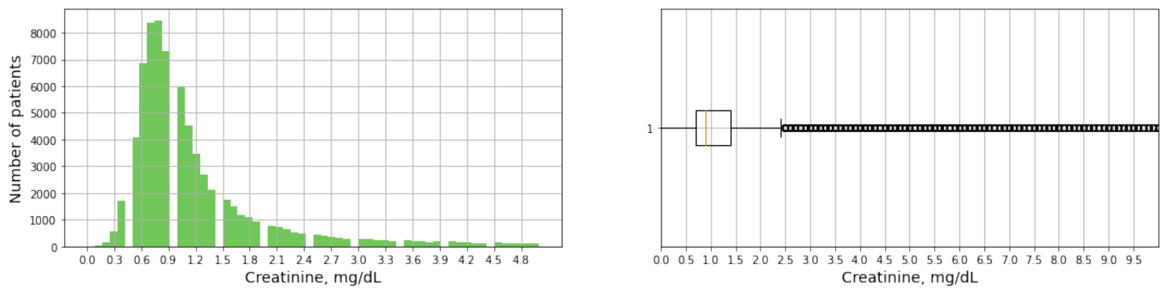
Table 3.1: Final cohort statistics description for LSTM-based and XGBoost models.

ratio of female patients to male patients is around 4 : 5. The median age is 65, the youngest patient is 18 years old, and the oldest patient is 100 years old. The most common race among all patients in the final cohort is white, followed by unknown (or other, unable to obtain), and African American. The data preprocessing step for the XGBoost model differs from the one for the LSTM model. Due to that, some samples were excluded from the dataset before being fed to the XGBoost model. Table3.1 shows the difference between the total number of samples and the description of the final cohort for both models.

The distribution of patients' weight is shown on figure 3.3. We can observe that the second quartile (i.e., median) is slightly higher for higher stages of AKI, e.g., for patients who did not develop an AKI onset, the median weight is 160 kg, for those who developed stage 1 of AKI it is 168 kg. In contrast, for those who developed stage 3 of AKI, the median weight is 172 kg. The first quartile follows a similar pattern.



(a) Distribution of patients having different lengths of stay in ICU.



(b) Distribution of patients having different creatinine levels on the first 24h in the ICU.

Figure 3.5: Distribution of patients having different lengths of ICU stay (a) and different creatinine levels on the first 24h in the ICU (b).

3.1.4 Exclusion criteria

In our problem setting, a sample is the records of a single ICU stay of a patient. In other words, we can have several samples corresponding to the same patient. Moreover, one patient can have several hospital admissions, and multiple ICU stays during that admission. MIMIC IV contains records of 76,943 unique ICU stays in total, which corresponds to 69,639 hospital admissions and 53,569 patients. If a sample falls into one of the following categories, we exclude it from the cohort:

- less than two creatinine measurements recorded;
- the first creatinine measurement is abnormal;
- length of stay in the ICU is less than two days;
- AKI of stage 2 or 3 occurred during the first 24h in the ICU.

We compared the obtained labels with diagnoses information from MIMIC IV. Since the diagnoses were entered into the database during patient discharge from the hospital, we have the diagnoses information for each admission without corresponding timestamps. Suppose the algorithm described above has detected AKI of any stage for a patient, and we have any AKI diagnoses (Table 3.2) recorded on the discharge for this patient. In that case, we mark this sample as True Positive (TP). Otherwise, we mark this sample as False Positive (FP). On the other hand, we marked as True Negative (TN) a sample

ICD version	ICD code	Diagnosis
9	5845	Acute kidney failure with lesion of tubular necrosis
9	5846	Acute kidney failure with lesion of renal cortical necrosis
9	5847	Acute kidney failure with lesion of renal medullary (papillary) necrosis
9	5848	Acute kidney failure with other specified pathological lesion in kidney
9	5849	Acute kidney failure, unspecified
10	N17	Acute kidney failure
10	N170	Acute kidney failure with tubular necrosis
10	N171	Acute kidney failure with acute cortical necrosis
10	N172	Acute kidney failure with medullary necrosis
10	N178	Other acute kidney failure
10	N179	Acute kidney failure, unspecified
10	N990	Postprocedural (acute) (chronic) kidney failure
10	O904	Postpartum acute kidney failure

Table 3.2: Diagnoses and corresponding ICD codes used to compare labels obtained using KDIGO criteria [22] and diagnoses records from MIMIC IV[21].

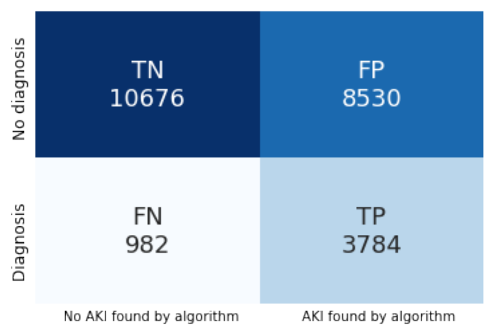


Figure 3.6: Confusion matrix describing the relationships between labels assigned by the AKI detecting algorithm and diagnoses records from MIMIC IV. *(Notice that samples with AKI of stage 2 or 3 occurred during the first 24h in the ICU are not excluded on this diagram.)*

corresponding to a patient without AKI detected by the algorithm and no AKI diagnoses found in the records. If there are any AKI diagnoses in the records, we mark this sample as False Negative (FN). Figure 3.6 shows the number of samples in each category.

The number of FP labels is relatively high, indicating many AKI cases have not been properly documented and entered in the table. However, we exclude the ambiguous cases (FP and FN) from the cohort to avoid potential noise.

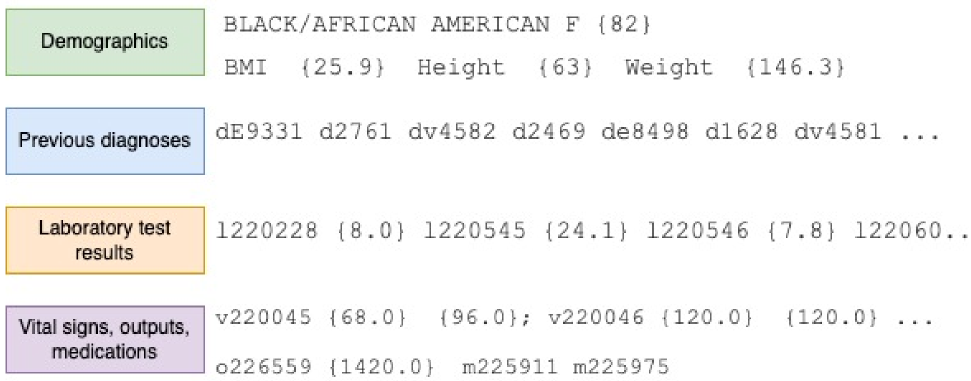


Figure 3.7: example of the input data passed to the model.

3.1.5 Predictor variables and input data

For such a fine-grained problem as predicting an AKI onset and its stage on a particular day of the ICU stay, it is crucial to use granular data. The input data is divided into two categories: *static* and *continuous*. The static data includes information that is not changing during the patient's stay, namely, demographics, diagnoses from previous patient admissions, and body measurements made before the transfer to ICU. On the other hand, the continuous data is changing with every new sample, particularly laboratory test results, vital signs, and volume of urine output. For the continuous values, we used maximum values during 24h period (for laboratory test results), minimum and maximum values for 12h period (vitals), and the total volume of urine output for 12h period. We also define taken medications as a continuous variable since they can change every time. Figure 3.7 shows the example of data blocks passed to the model. The paragraphs below describe each of the blocks.

Demographics and body measurements The demographics block consists of *age, gender* and *race* information. Besides, we added other semi-static body measurements recorded by the Online Medical Record system before the ICU stay, such as *body mass index (BMI), weight, heights, blood pressure*. If there are available records in the last 365 days, we take the latest ones. Otherwise, we do not use this information because these parameters could change drastically from when they were recorded and potentially infuse unwanted noise.

Previous diagnoses In this block, we included all diseases diagnosed in the patient for all time, except the current admission. There are 13,800 unique diagnoses ICD codes that were used as input data for the model. Each code was converted into a textual format and concatenated to the *'d'* letter.

Laboratory tests This block contains all laboratory test results made during the observation window of the ICU stay for each 24h period. If more than one test was made, we used the maximum value. Every laboratory test was encoded using the MIMIC IV original *itemid* number and concatenated to the *'l'* letter. For example, *Potassium (serum)* encoded as *l227442*. Each code is followed by a number, representing the test results and converted to a textual format. The results of the following laboratory tests related to AKI (according to Chen et al. [8]) were used as predictor variables: *anion gap, albumin, bands, bilirubin, hematocrit, lactate, sodium, bicarbonate, blood urea nitrogen (BUN), calcium, chloride, creatinine, hemoglobin, international normalized ratio (INR), platelet, potassium, prothrombin time (PT), partial thromboplastin time (PPT), white blood cell count (WBC), and glucose.*

Vital signs, outputs, and medications The current block contains vital signs, urine outputs, and given medications for each 12h period during the observation time. Similarly to laboratory test results, each type of measurement was assigned a unique code, corresponding to its *itemid* in MIMIC IV, and concatenated with the letters *v*, *o*, *m* for vital sign, urine output, and medication correspondingly. Since vital signs information is collected with the help of *Metavision* bedside monitors, many measurements are recorded for each day. Therefore, only maximum and minimum values for each 12h period were used in the model. The urine output value is collected approximately every two hours, so the sum of the urine output is calculated for 12h (in milliliters). Urine output coming from different sources (e.g., foley or condom catheter) have different *itemid* and, consequently, different codes. For medications, we used corresponding *itemid* to code each medication given during 12h period.

3.2 Prediction model

3.2.1 LSTM structure

Long Short-Time Memory is one of the types of RNN which alleviates the typical for RNNs problem of vanishing (exploding) gradients. It was introduced in 1997 by Hochreiter, and Schmidhuber [18]. This model allows using long-term dependencies of the input data with the help of three *gate layers*. Figure 3.8 shows the structure of the single LSTM cell, and the equations below show the calculation process inside this cell, where f_t, u_t, and o_t are *forget, update* and *output* gate layers correspondingly.

Forget gate (eq. 3.1) helps the model to decide how much information to keep from the previous time step. It is calculated by first multiplying the x_t (input) and h_{t-1} (the cell state from the previous time step) by weight matrices of input U_f and W_f, then summing up the multiplication results and adding the bias b_f. Then, the sigmoid function σ is applied to get a value ranging from 0 to 1. The closer the result to 0, the more information will be forgotten, while closeness to 1 means keeping most of the information.

Update gate (eq. 3.2) is responsible for deciding which information from the input should be stored in the current cell state. Similarly to the forget gate, it is calculated by summing up the bias b_u, the product of input x_t and weight matrix U_u, the product of

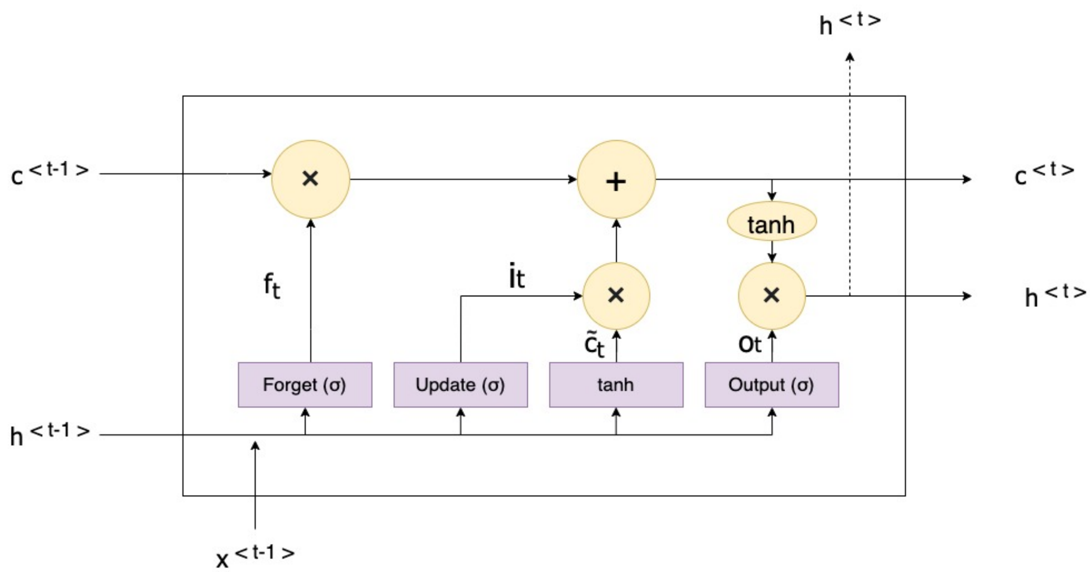


Figure 3.8: The structure of a single LSTM cell.

the hidden state from the previous time step h_{t-1} and weight matrix W_u. Similarly, the sigmoid function is applied to the result. In addition, the vector with candidate values N_t is calculated using the tanh function and weight matrices U_n and W_n, as well as the bias b_n (eq. 3.4). Since the *tanh* function outputs the values from -1 to 1 it allows to add or subtract the candidate values.

Then, the previous cell state (eq. 3.5) c_{t-1} is updated to the new cell state c_t by adding the Hadamard product (element-wise multiplication) of c_{t-1} and f_t to the Hadamard product of u_t and N_t.

Output gate (eq. 3.3) decides what information from the cell state the model will output as the new hidden state. The output gate o_t is calculated similarly to f_t and u_t, but using the weight matrices U_t and W_t, as well as the bias b_o. Then, the hidden state (eq. 3.6) h_t is calculated by applying the *tanh* function over the new cell state c_t and multiplying element-wise the o_t with the result.

In addition to the new input, the new cell state and hidden state are passed as the input to the next cell of the LSTM model.

$$f_t = \sigma(x_t U_f + h_{t-1} W_f + b_f) \tag{3.1}$$

$$u_t = \sigma(x_t U_u + h_{t-1} W_u + b_u) \tag{3.2}$$

$$o_t = \sigma(x_t U_o + h_{t-1} W_o + b_o)) \tag{3.3}$$

$$N_t = \tanh(x_t U_n + h_{t-1} W_n + b_n) \tag{3.4}$$

$$C_t = f_t * C_{t-1} + u_t * N_t \tag{3.5}$$

$$h_t = o_t * \tanh(C_t) \tag{3.6}$$

3.2.2 Architecture of the proposed model

The model is based on the LSTM neural network since it allows to capture long term dependencies in the input data. Due to its hight length, the input data, described in section 3.1.5 are combined into four blocks:

- demographics (which includes both: demographic information and semi-static body measurements);
- previous diagnoses;
- laboratory test results taken during the first 24h in the ICU;
- vital signs, urine output volume recorded during two 12h windows and medications prescribed during the same two 12h windows from the time of transferring to the ICU.

Every number in the text was enclosed in brackets in order to pass the sense of the start and the end of the numerical information to the model. Each of these blocks is converted into a tensor and then tokenized using Byte-Pair Encoding (BPE) technique. For each of the tensors, we define the *maximum length* in tokens. Thus, if an input text is tokenized into more tokens than the *maximum length*, we truncate it. On the other hand, if the text input has fewer tokens than the *maximum length*, we pad the sequence with special token *'PAD'*. Each tokenized block has a special token at the beginning and end of the text (before the padding tokens).

Then, we concatenated the static information blocks - demographics and previous diagnoses. Each resulting tokenized vector is passed to an embedding layer to get the representations of the input text. The representations of the static information are passed to a fully-connected layer, while the representations of the continuous information are passed to the LSTM layer. After that, the outputs of both layers are concatenated and passed to the other fully-connected layer. Then the output is passed to the bidirectional LSTM layer. Finally, the dropout is applied to the output from the LSTM layer, and then the resulting tensor is passed through another fully-connected layer with three output nodes.

Then, the sigmoid activation function is applied to the model's output. Here we defined the problem of predicting the AKI and its stage as the multi-label classification.

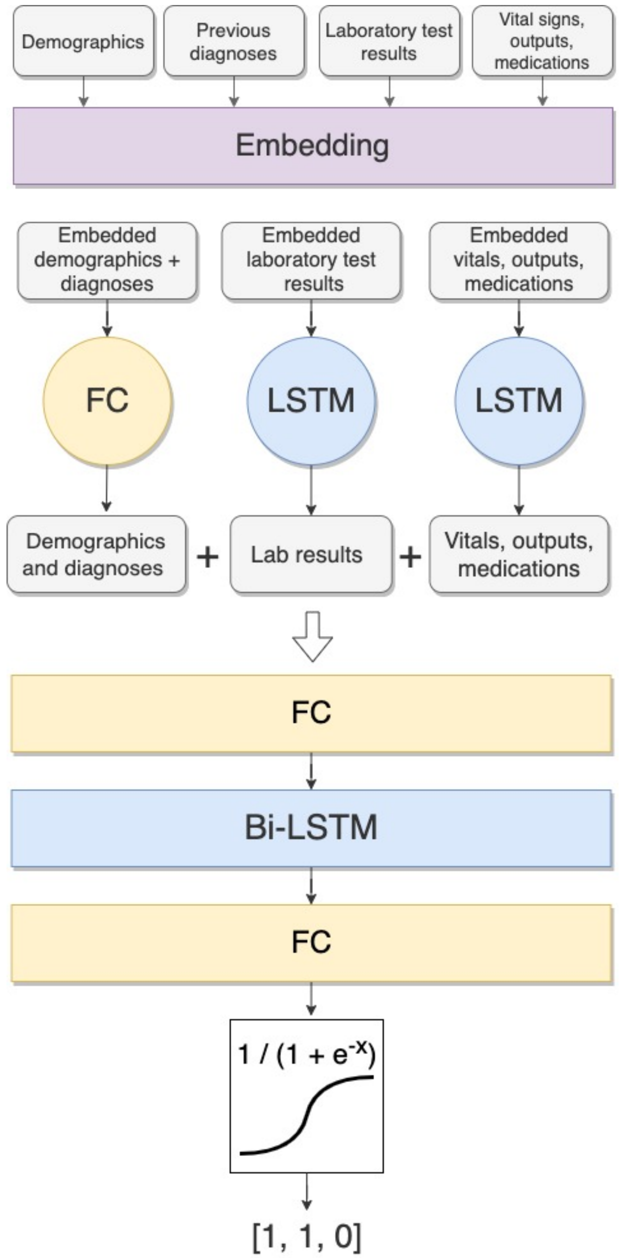


Figure 3.9: The architecture of the proposed model.

3.2.3 XGBoost AKI prediction model

Extreme Gradient Boosting or XGBoost [7] is a machine learning technique based on gradient boosting decision trees (GBDT). The gradient boosting is an ensemble of shallow decision trees where the error residuals from the previous trees are used to train the next set of models. The training process is wrapped into the gradient descent algorithm over an objective function.

Before feeding data into the XGBoost model, it needed to be preprocessed in a certain way. The demographic information consisted of BMI, height, weight, diastolic and systolic

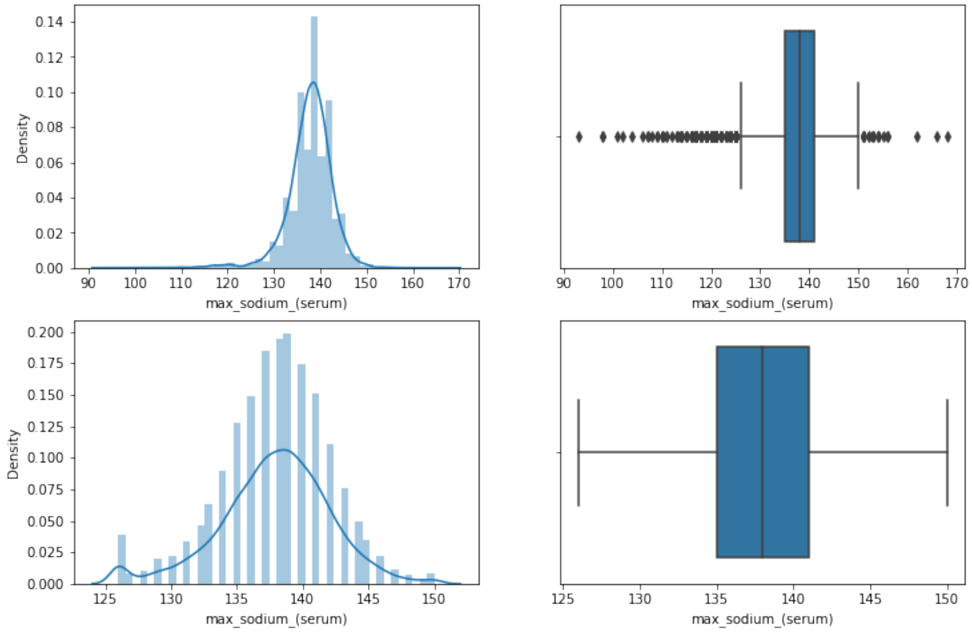


Figure 3.10: The distribution of the sodium in the training cohort before and after the data normalization using the method described in equations 3.7 and 3.8.

blood pressure (the latest value recorded within 365 days before admission into the ICU), as well as age, gender, and race. The minimum and maximum vital signs values for 24h were used: heart rate, blood pressure measurements, and body temperature. Similarly to vital sights, the minimum and maximum values of laboratory test results were used as the input features. The sum of urine output for 24h was also the input feature.

Previous diagnoses were coded as 1 if a patient has been diagnosed with the disease earlier in their life. Otherwise, it was coded as 0. The set of diagnoses was chosen according to the analysis of the impact of different features on the risk of AKI conducted by Chen et al. [8]. We used the following diagnoses: *cirrosis, congestive heart failure (CHF), coronary artery disease (cad), liver disease, myocardial infarction (MI), diabetes, hypertension, peripheral vascular (PV).*

The set of medications was also chosen with the help of the study of Chen et al. [8]. Since the medications in this study are represented as Anatomical Therapeutic Chemical (ATC) Classification codes, and the medications in MIMIC-IV are represented as National Drug Code (NDC), we needed to map the ATC codes into the NCD codes. The mapping was performed using the database from the original US Food and Drug Administration website [3].

We imputed the missing values of the continuous variables with the average value in the cohort. After that, the data was normalized using the interquartile method described in the equations 3.7 and 3.8, where x is the target value, $Q1$ is the first quartile (a value higher or equal to 25% of all data points), and $Q3$ is the third quartile (a value higher or equal to 75% of all data points).

[3] `https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory`

$$
\begin{aligned}
IQR &= Q3 - Q1 \\
U &= Q3 + 1.5 * IQR \\
L &= Q1 - 1.5 * IQR
\end{aligned}
\tag{3.7}
$$

$$
\begin{aligned}
\text{if} \quad & x \geq U, \quad x = U \\
\text{if} \quad & x \leq L, \quad x = L \\
\text{else} \quad & x = x
\end{aligned}
\tag{3.8}
$$

Figure 3.10 shows the distribution of sodium in the training cohort before and after the data normalization.

The training, testing, and validation data splits were imputed and normalized separately to prevent information leakage from training data into testing data.

Chapter 4

Experiments and Results

4.1 XGBoost models

4.1.1 Experimental design

The baseline models were implemented using python XGBoost package [1]. All models had 500 estimators and eta of 0.3. Parameters such as 'max_depth", "min_child_weight", "colsample_bytree", and "subsample" were tuned using grid search and 5-fold cross-validation. In each experiment, three models were trained with the same data but with different sets of labels corresponding to each subtask. Therefore, three sets of parameters were found for each model (Table 4.1).

	Subtask #1	**Subtask #2**	**Subtask #3**
max_depth	3	3	3
min_child_weight	7	7	9
subsample	1	1	1
colsample	1	1	1

Table 4.1: Parameters found using grid serach and 5-fold cross-validation for each set of labels.

To tackle the problem of high imbalance in the dataset, we experimented with several techniques and their combinations: oversampling, undersampling, and weighted XGBoost. The positive weights for each experiment were calculated using the following formula: $pos\_weight = \frac{n}{n\_pos}$, where n is a number of samples in dataset, and n_pos is the number of positive samples.

The sampling strategy was defined as follows: for subtasks #1 and #2, the ratio of positives to negatives after oversampling is equal to 0.5, and after undersampling, it is equal to 0.4; and for subtask #3, the ratio of positives to negatives after oversampling

[1] `https://xgboost.ai/`

Subtask #1			
	n_neg	n_pos	pos/neg
Original	9293	1194	0.13
Oversampling	9293	4646	0.5
Undersampling	2985	1194	0.4

Subtask #2			
	n_neg	n_pos	pos/neg
Original	10056	431	0.05
Oversampling	10056	5028	0.5
Undersampling	1077	431	0.4

Subtask #3			
	n_neg	n_pos	pos/neg
Original	10401	86	0.008
Oversampling	10401	2080	0.2
Undersampling	430	86	0.2

Figure 4.1: The quantitative description of the training data fed to XGBoost algorithm, before and after applying sampling techniques.

is equal to 0.2, and after undersampling it is equal to 0.2 as well. Figure 4.1 shows the oversampling and undersampling strategy used to handle an imbalance in the data.

For the experiments where we combine the weighted XGBoost and oversampling or undersampling, we first sampled the data and then calculated weights for the given experiment. As a result, the experiments with original and sampled data have different weights.

4.1.2 Results

Since the task of predicting an AKI onset consisted of three separate subtasks with different sets of labels, the performance is shown for all three subtasks.

Table 4.2 shows F1-score, sensitivity, and precision scores for each of the experiments described in the Section 4.1.1. The experiments were conducted for each set of labels separately: for any stage of AKI, stages 2 or 3, and stages 3 of AKI. The sets of labels correspond to subtasks #1, #2, and #3, respectively. To get the prediction, we compared the probability value outputted by the model. If the value was higher or equal to the chosen threshold, the sample was labeled 1, otherwise 0. We defined a set of candidate thresholds to choose the threshold, then calculated the predictions based on it. Using the resulting predictions, we computed F1 scores for each of the thresholds from the set. The results presented in these tables are shown for the threshold, which gives the highest F1 score. Table 4.3 shows other metrics describing the performance of the models in the same experiments. The primary metrics here are PR AUC and F1-score since the data is highly imbalanced, and the problem of predicting AKI puts more importance into predicting the positive class correctly. Another critical metric is 33% precision - sensitivity, which shows the model's sensitivity at the minimum acceptable precision threshold. We chose this precision threshold based on the study of Tomasev et al.[40].

Subtask #1				
Experiment	**F1-score**	**Sensitivity**	**Precision**	**TP:FP**
XGB	0.51	0.55	0.48	1.1
Weighted XGB	0.49	0.71	0.37	1.7
XGB & oversampling	0.53	0.59	0.48	1.1
XGB & undersampling	0.48	0.71	0.37	1.8
Weighted XGB & oversampling	**0.55**	0.54	0.57	0.8
Weighted XGB & undersampling	0.54	0.69	0.44	1.3
Subtask #2				
XGB	0.34	0.5	0.25	3
Weighted XGB	0.28	0.7	0.18	4.7
XGB & oversampling	0.3	0.79	0.19	4.4
XGB & undersampling	0.37	0.59	0.27	2.8
Weighted XGB & oversampling	0.35	0.52	0.27	2.8
Weighted XGB & undersampling	**0.4**	0.46	0.35	1.9
Subtask #3				
XGB	**0.21**	0.2	0.23	3.4
Weighted XGB	0.12	0.13	0.12	7.5
XGB & oversampling	**0.21**	0.2	0.23	3.4
XGB & undersampling	0.18	0.06	0.11	8.3
Weighted XGB & oversampling	0.2	0.33	0.15	5.8
Weighted XGB & undersampling	0.19	0.53	0.12	7.5

Table 4.2: Results for experiments with XGBoost (XGB) model and different balancing techniques for predicting if a patient will develop AKI during the second day in the ICU. The last column corresponds to the amount of false positive predictions for each true positive.

From the results corresponding to subtask #1, we can see that the weighted XGBoost model combined with undersampling and the weighted XGBoost model combined with oversampling techniques achieved better F1-scores (0.55 and 0.54) compared to other experiments (table 4.2). However, the sensitivity of weighted XGBoost with undersampling is much higher than in weighted XGBoost with oversampling. Moreover, when we consider 33% precision as a threshold to increase the sensitivity of the model while keeping the precision on the minimum reasonable level, weighted XGBoost with undersampling achieves the performance of 0.79 (table 4.3). This means that this model is capable of detecting more patients at risk of developing AKI. From figure 4.2a, we can observe that this model focused on the following features the most: creatinine, INR, hemoglobin, BUN, diuretics drugs (C03), and patients' weight.

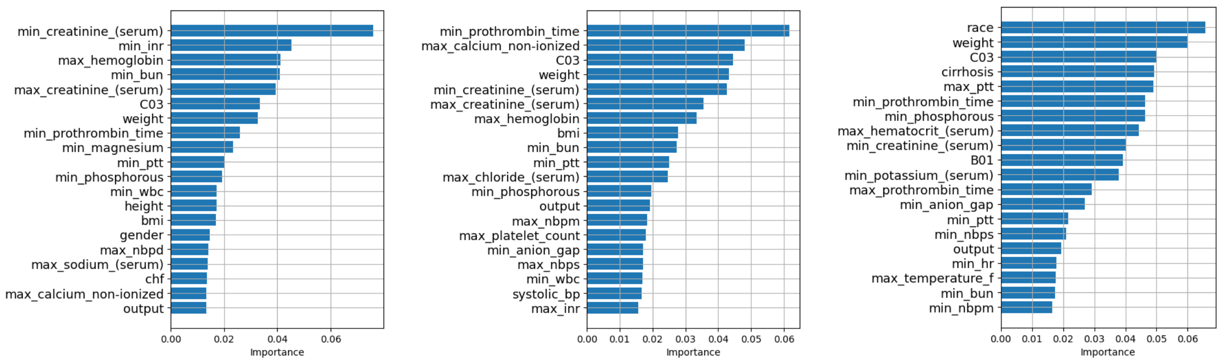
Regarding subtask #2, we can observe that the weighted model with undersampling achieved the best F1-score (0.4) among all other experiments, while the sensitivity of this model is the lowest. The second best model in terms of F1-score in the current subtask is XGBoost with undersampling. However, the precision of this model is much lower. These two models also achieved the highest sensitivity given the 33% precision threshold, while

Subtask #1			
Experiment	**PR AUC**	**ROC AUC**	**33% Precision Sensitivity**
XGB	0.53	0.74	0.74
Weighted XGB	0.54	0.78	0.75
XGB & oversampling	0.54	0.75	0.77
XGB & undersampling	0.49	0.78	0.73
Weighted XGB & oversampling	**0.56**	0.74	0.75
Weighted XGB & undersampling	0.51	0.79	**0.79**
Subtask #2			
XGB	0.27	0.72	0.16
Weighted XGB	0.18	0.78	0.07
XGB & oversampling	0.21	0.82	0.09
XGB & undersampling	0.25	0.76	0.39
Weighted XGB & oversampling	0.24	0.73	0.16
Weighted XGB & undersampling	**0.33**	0.71	**0.46**
Subtask #3			
XGB	0.07	0.6	0.07
Weighted XGB	0.07	0.56	-
XGB & oversampling	0.07	0.6	0.07
XGB & undersampling	0.08	0.77	-
Weighted XGB & oversampling	0.08	0.66	-
Weighted XGB & undersampling	0.08	0.74	0.07

Table 4.3: Additional results of experiments with XGBoost (XGB). The metrics go as follows: precision-recall area under the curve (PR AUC), receiver operating characteristic curve (ROC AUC), and the value for sensitivity when the threshold is chosen for the precision to be equal to 33%.

other models failed to provide a reasonable level of sensitivity at this threshold. This model puts more weight on the following features: prothrombin time, calcium, diuretics drugs, patients' weight, and creatinine 4.2b. The model with oversampling yielded the best sensitivity (0.79). However, the precision score is very low in this experiment, indicating the poor ability of the model to discriminate between positive and negative classes and tends to assign positive labels to negative examples.

In subtask #3, the original XGBoost model alone and combined with oversampling technique achieved better F1 scores than other experiments. Nevertheless, the sensitivity in both experiments is very low and does not provide the required efficiency. In this subtask, none of the balancing techniques improved the actual performance of the model summarized in the F1-score metric and 33% precision - sensitivity, neither class weights nor sampling techniques. However, we can see that the features' importance in this model is distributed differently from the models in the other two subtasks. The most important features are race, weight, diuretics, the presence of cirrhosis diagnosis in a patient's history, and PTT 4.2c.



(a) Subtask #1 (b) Subtask #2 (c) Subtask #3

Figure 4.2: Top features contributed to the decision of weighted XGBoost model with an undersampled dataset for each subtask.

4.2 LSTM-based model

4.2.1 Experimental design

The model was implemented using the Pytorch library [31]. The Adam optimization algorithm [24] was used during the training process. The loss function was the Binary Cross Entropy, shown in the equation 4.1, where $n \in \{0, \ldots, N\}$ and N is the number of training samples in a batch. The learning rate was set to 0.00001, and the vocabulary size was 8791. The hidden size for LSTM layers was 128. All results are reported as an average and standard deviation of the performance for three identical experiments.

$$
\begin{aligned}
l_n &= -w_n[y_n \log x_n + (1 - y_n) \log(1 - x_n)] \\
\mathcal{L}(x, y) &= mean(\{l_1, \ldots, l_n\})
\end{aligned}
\tag{4.1}
$$

The model was fed data blocks described in the section 3.1.5 containing the information of a patient recorded during the first 24h in the ICU ward, i.e., the observation period is equal to 24h. The model target is to predict if a patient will develop AKI of a given stage during the next 24h after the observation period, i.e., the prediction period is equal to 24h. Figure 4.3 shows the described settings.

Each data block was coded into a set of unique token identifiers using Byte-Pair encoding tokenization technique (BPE) implemented with the help of the *hugging face* library [2]. The output vector of the model was activated using the sigmoid function 4.2.

$$
S(x) = \frac{1}{1 + e^{-x}} \tag{4.2}
$$

After applying the activation function, the threshold was chosen using the same method as in the baseline experiments with XGBoost (Section 4.1.2), and the predictions were made based on this threshold.

[2] `https://huggingface.co/`

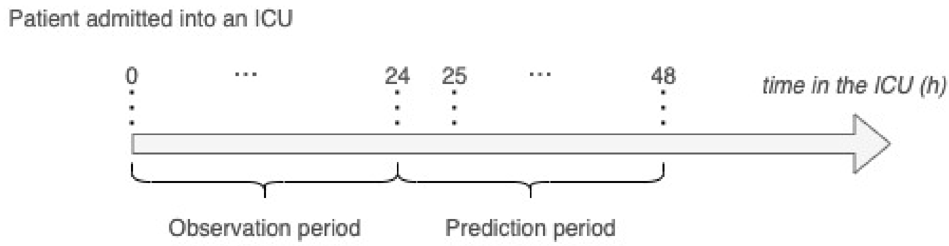


Figure 4.3: The experimental design for defining the observation and prediction periods. The medical records from the first 24h in the ICU are used as input for the models. The second 24h are used to define the labels: if a patient developed AKI onset of target stage during this period - it assigned label 1, otherwise 0.

4.2.2 Results

The results in tables 4.4 and 4.5 show that within the scope of the subtask #1, the model can learn some useful dependencies between the data. The best F1-score of 0.39 was achieved by the model with a dropout probability of 0.6 and an embedding size of 150. The same model also achieved the best specificity, meaning that it can detect the negative samples more often and the best precision, which implies that the classification of positive samples is more accurate than in other experiments. The latter can also be seen from the true positive ratio in the table 4.5 - the same model had the lowest false positive predictions for each true positive. The other model with the same embedding size but a lower dropout of 0.4 achieved the best sensitivity and the second-best F1 score. Overall, the models with smaller embedding sizes achieved better PR AUC and sensitivity with a precision threshold of 33%. *Note that here the precision threshold is described, not the probability threshold.*

For the subtask #2 experiments, we can see that the models still learned some valuable dependencies. The models with an embedding size of 200 and dropout probability of 0.6, and vice versa, yielded better F1-score than models with other settings. However, table 4.5 shows that no models could pass the given precision threshold of 33% and achieve any reasonable sensitivity score.

Similarly, the models trained using the labels for stage 3 of AKI within the scope of the subtask #3 also could not pass the given threshold. In this setting, all models failed to learn useful information to make reasonable predictions, which can be seen from the low values of the F1-score and PR AUC. Even though the values of ROC AUC are relatively high, since the dataset is highly imbalanced, it does not show the model's actual performance.

Model interpretation

Since the proposed model is based on LSTM neural network, which is hard to interpret, the Local Interpretable Model-agnostic Explanations (LIME) [36] algorithm was used to visualize the model's explanations. LIME interprets the features used by the model to make predictions by local linear approximation. The algorithm implemented by *lime* library[3] was

[3] `https://github.com/marcotcr/lime`

Subtask #1						
Exp.	**Dropout**	**Emb. size**	**F1-score**	**Sensitivity**	**Precision**	**TP:FP**
1	0.1	200	0.36±0.01	0.61±0.06	0.26±0.02	2.9
2	0.1	150	0.37±0.01	**0.64±0.05**	0.26±0.02	2.9
3	0.2	200	0.37±0.01	0.58±0.09	0.27±0.03	2.7
4	0.2	150	0.37±0.01	0.54±0.06	0.28±0.02	2.6
5	0.4	200	0.33±0.02	0.51±0.08	0.24±0.03	3.2
6	0.4	150	0.37±0.01	0.61±0.05	0.27±0.01	2.8
7	0.6	200	0.36±0.04	0.53±0.02	0.27±0.04	2.8
8	0.6	150	**0.39±0.01**	0.57±0.05	**0.29±0.01**	**2.5**
Subtask #2						
Exp.	**Dropout**	**Emb. size**	**F1-score**	**Sensitivity**	**Precision**	**TP:FP**
1	0.1	200	0.19±0.02	0.36±0.06	0.13±0.01	6.8
2	0.1	150	0.19±0.01	0.44±0.04	0.12±0.01	7.0
3	0.2	200	**0.22±0.02**	0.38±0.02	**0.15±0.02**	**5.6**
4	0.2	150	0.18±0.01	0.37±0.08	0.13±0.01	7.0
5	0.4	200	0.14±0.01	**0.46±0.12**	0.09±0.02	10.8
6	0.4	150	0.18±0.01	0.4±0.14	0.12±0.01	7.6
7	0.6	200	0.18±0.03	0.42±0.12	0.12±0.03	7.93
8	0.6	150	0.17±0.01	**0.46±0.04**	0.11±0.01	8.4
Subtask #3						
Exp.	**Dropout**	**Emb. size**	**F1-score**	**Sensitivity**	**Precision**	**TP:FP**
1	0.1	200	0.05±0.01	0.52±0.32	0.03±0.01	40.0
2	0.1	150	0.05±0.03	0.33±0.35	0.03±0.02	38.3
3	0.2	200	0.02±0.02	0.27±0.4	0.02±0.02	49.0
4	0.2	150	0.02±0.02	0.3±0.45	0.01±0.01	37.5
5	0.4	200	0.01±0.01	0.73±0.28	0.0±0.01	169.67
6	0.4	150	0.03±0.01	0.64±0.26	0.02±0.01	90.9
7	0.6	200	0.02±0.02	0.49±0.37	0.02±0.02	126.67
8	0.6	150	0.02±0.0	0.76±0.19	0.01±0.0	117.23

Table 4.4: Results for experiments with different dropout probabilities and embedding size of LSTM-based model for predicting if a patient will develop AKI during the second day in the ICU. The last column corresponds to the amount of false positive predictions for each true positive.

used in this work.

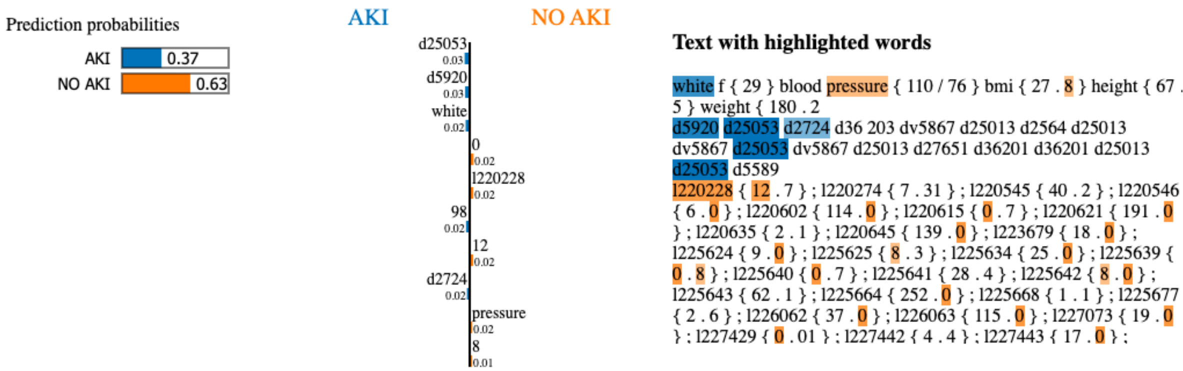
Figure 4.4 shows the explanations of the model from experiment 2 for testing samples of subtask #1. In both examples, the model paid much attention to zeroes without their context as the predictors of a negative outcome (NO AKI), which is incorrect. This probably contributed to the final prediction made by the model. However, figure 4.4a the model correctly recognized the presence of diagnoses *5920 - calculus of kidney (i.e., kidney stones, or nephrolithiasis) and 25053 - the type of diabetes* as a predictor of a positive outcome (AKI). The multiple studies showed that these commodities are associated with

Subtask #1					
Exp.	**Dropout**	**Emb. size**	**PR AUC**	**ROC AUC**	**33% Precision Sensitivity**
1	0.1	200	0.3±0.01	0.71±0.01	0.38±0.03
2	0.1	150	0.3±0.01	0.72±0.01	0.37±0.03
3	0.2	200	0.3±0.01	0.71±0.02	0.38±0.04
4	0.2	150	**0.31±0.02**	0.7±0.01	0.38±0.02
5	0.4	200	0.23±0.03	0.67±0.02	0.11±0.17
6	0.4	150	0.3±0.01	**0.71±0.01**	**0.43±0.01**
7	0.6	200	0.27±0.04	0.69±0.03	0.33±0.08
8	0.6	150	0.3±0.01	**0.71±0.02**	0.38±0.11
Subtask #2					
Exp.	**Dropout**	**Emb. size**	**PR AUC**	**ROC AUC**	**33% Precision Sensitivity**
1	0.1	200	0.09±0.01	0.64±0.03	-
2	0.1	150	0.09±0.01	**0.67±0.01**	0.02±0.0
3	0.2	200	0.09±0.01	0.65±0.01	-
4	0.2	150	0.09±0.01	0.64±0.03	-
5	0.4	200	0.07±0.01	0.64±0.03	-
6	0.4	150	0.09±0.01	0.65±0.04	-
7	0.6	200	0.08±0.01	0.65±0.03	-
8	0.6	150	**0.10±0.01**	**0.67±0.04**	0.07±0.0
Subtask #3					
Exp.	**Dropout**	**Emb. size**	**PR AUC**	**ROC AUC**	**33% Precision Sensitivity**
1	0.1	200	0.02±0.01	0.67±0.1	-
2	0.1	150	0.02±0.0	0.58±0.06	-
3	0.2	200	0.02±0.0	0.53±0.11	-
4	0.2	150	0.02±0.01	0.59±0.14	-
5	0.4	200	0.0±0.01	0.42±0.08	-
6	0.4	150	0.02±0.01	0.57±0.09	-
7	0.6	200	0.01±0.01	0.44±0.11	-
8	0.6	150	0.01±0.0	0.53±0.04	-

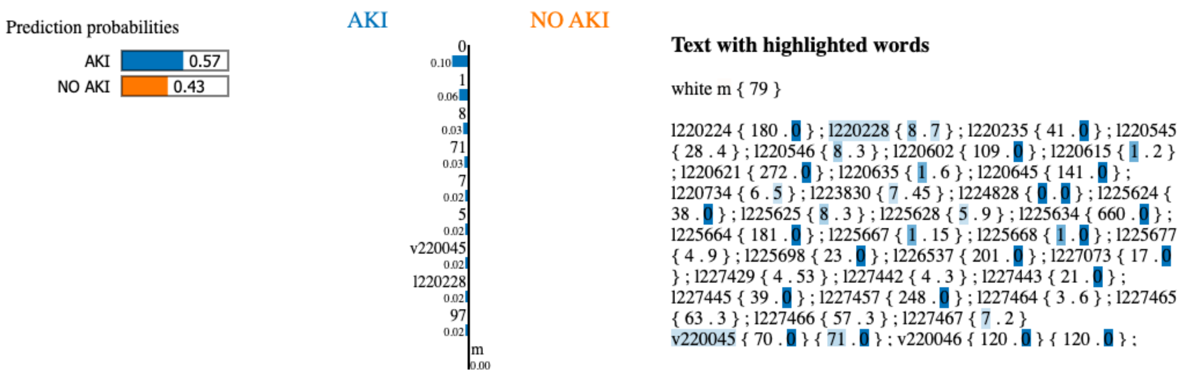
Table 4.5: Other metrics describing the results of experiments with LSTM-based model for predicting if a patient will develop AKI during the second day in the ICU. The first two columns describe the dropout probabilities and embedding size of the model.

the increased risk of developing AKI [16][32][39].

The model also identified the hemoglobin laboratory test result coded as *l220228* as meaningful for predicting AKI. The study of Seung Seok Han [17] showed that the patients with anemia (the hemoglobin is < 10.5 g/dL) are at a higher risk of AKI compared to ones with normal hemoglobin level (≥ 10.5 g/dL). The explanation of the true negative example indicates that the model used the normal hemoglobin level (12.7 g/dL) as a negative



(a) True negative example



(b) True positive example

Figure 4.4: Local interpretations of true negative (a) and true positive (b) examples from the testing dataset for LSTM-based model in experiment 2.

predictor of AKI. In contrast, the explanation of the true positive example suggests that the patient with anemia (hemoglobin is 8.7 g/dL) is at a higher risk of developing AKI. Figure 4.4b also shows that the model used the maximum heart rate of a patient (*code v220045*) as a contributor to the risk of developing AKI.

Chapter 5

Discussion

This work aims to tackle the following research objectives: *1) developing a labeling algorithm for AKI predicting task; 2) developing a machine learning framework to predict AKI on the second day of ICU stay; 3) exploring and evaluating different approaches to alleviate the problem of high imbalance in the data; and 4) exploring the use of LSTM-based approach to a problem of a daily AKI prediction.* In each experiment, we predicted the AKI outcome on the second day in the ICU from the medical records made on the first day. Each experiment was evaluated according to three subtasks: *1) predicting any stage AKI onset; 2) predicting the AKI onset of stage 2 or 3; 3) predicting the stage 3 AKI onset.* In the first two subtasks, the results of the experiments suggest that utilizing the undersampling technique is more efficient in enhancing the model performance in the data imbalance settings. It is consistent with the study of Yuan Wang et al. [42], where they showed that the undersampling technique helps to tackle the imbalance problem better, compared to oversampling. The XGBoost with undersampling dataset and weights showed reasonable performance comparable to other studies in the field [40][42] in the first subtask. The experiments in the third subtask did not yield clinically applicable performance. In general, the XGBoost model had better performance than the LSTM-based model.

The labeling algorithm. The labeling algorithm proposed in this work was able to obtain the labels for each day of the patient's ICU stay using the creatinine laboratory test results and urine output records. However, the algorithm labeled as having AKI some samples corresponding to patients who were not diagnosed with AKI during the given admission 3.6. It was confirmed by the practicing nephrologist that the satisfaction of the AKI criteria is the only correct way to diagnose AKI. Therefore, the 'false positive' labels may be caused by the fact that clinicians did not document the AKI case on the discharge. The other possible reason may be the absence of some urine output records. For example, if a patient had a normal urine output rate that does not satisfy any AKI stage criteria, but some amount of the urine output was not recorded, it may lead the algorithm to assign the positive label based on the urine output part of the KDIGO[22] criteria. Similarly, for the negative labels, assigned to patients who had any of the AKI diagnoses 3.2 in the discharge summary. The negative label indicates that there were no abnormalities in creatinine satisfying the AKI criteria. However, the corresponding laboratory test result may not be recorded in the system, which leads the algorithm to a wrong decision. Another possible reason of the inconsistency is a baseline creatinine definition. Suppose a patient did not have any previous records before ICU admission and was not excluded from the cohort based on abnormal creatinine from the first laboratory test. Due to the absence of the baseline, the algorithm may assign the negative label, even though this patient developed AKI. Similarly, if a patient underwent the last laboratory test for creatinine earlier than seven days from the given day, the median value of all previous creatinine measurements for the previous year is taken as a baseline, which may be misleading and not represent the actual patient's creatinine baseline.

XGBoost model and data balancing. In the experiments corresponding to subtask #1, the undersampling technique combined with the weighted model yielded the best results in detecting correctly maximum patients at risk of AKI with the given acceptable precision threshold. Therefore, if we aim to find as many patients at risk as possible, accepting the precision of 33% tradeoff - this model can be the most effective.

When it comes to subtask #2 (the model trained on a set of labels corresponding to stages 2 and 3), the performance of the model drops. The most likely reason behind that is the decreased ratio of positive samples in the data (from 0.13 to 0.05, see figure 4.1). Since the precision of the models in these settings is relatively low, it is reasonable to use 33%-precision sensitivity as the primary metric. In this setting, weighted XGBoost with undersampling achieved better performance, which corresponds with the results of subtask #1.

In subtask #3, since the ratio of positive samples in the data is very low for stage 3 of the AKI prediction task (0.008, see figure 4.1), all models failed to label the samples correctly in the majority of the cases. Thus, the balancing techniques did not improve the performance on the target metrics. In addition, the set of unique samples was the same in all three subtasks. Only the labels were changed depending on the subtask. The samples with the lower stage of AKI are considered negative in the current subtask, e.g., the sample corresponding to a patient having AKI of stage 2 onset will be treated as negative in the third subtask. It suggests that the model was encountering noisy labels if the difference between AKI stages is not significant in terms of produced medical records by a patient.

The models in subtasks #1 and #2 are paying more attention to the laboratory test results such as creatinine, INR, hemoglobin, and BUN, as well as to diuretics medications and patients' weight. On the contrary, the importance of analysis shows that the mode in subtask #3 focuses on race and weight the most and then on diuretics and cirrhosis diagnosis in the history of a patient. This indicates that patients with higher weight are prone to having more severe AKI courses than the ones with lower weight, which is supported by the study on the impact of obesity on the risk of AKI[11]. This is also supported by the fact that patients' median weight is higher in the patients who developed higher stages of AKI (Figure 3.3). The actual importance of race can be skewed because of the prevalence of *'white'* race in the data. The contribution of diuretics is also consistent with the study of Kipyo et al.[23].

Therefore, the XGBoost model in this setting can predict an AKI onset on the second day in the ICU, using the data from the first day in the ICU. However, the model fails when it comes to predicting stage 3 of AKI.

LSTL-based approach. For the daily AKI prediction problem wrapped into a text classification task, the pattern is similar to the XGBoost model. For the subtask of predicting the first stage of AKI (or any stage of AKI), the model was able to learn some useful information. However, it fails when the ratio of positive samples decreases. This model correctly identified the words denoting diabetes and kidney stone disease, as well as the low level of hemoglobin as the features associated with the higher risk of AKI. On the other hand, the model defined the normal level of hemoglobin as a predictor of a negative outcome (NO AKI).

The other interesting finding is that the models with an embedding size of 150 have better performance across all metrics in the first subtask. It indicates that a lower number of parameters positively affects the performance since the number of training samples is low, and it may be overwhelming for the model to learn low-level data features. Possibly, having more data could cause the opposite effect since, generally, more parameters help the model learn more high-level features of the sample. The higher dropout rates improved models' F1-score and sensitivity in subtask #1, while the lower rates improved the models on the same metrics in subtask #2.

Contributions and limitations. This work has several main contributions. *First*, the extensive evaluation of the baseline method XGBoost and experimenting with different balancing techniques was performed. *Second*, the problem of predicting AKI was tackled with three different granularity levels - any stage of AKI, stages 2 and 3, and stage 3 alone. *Third*, the results are reported with maximum transparency using metrics, which would be helpful to see the big picture of the models' performance. *Forth*, the transparency with the labels assigning process and the exclusion of the labels which are not fully clear, i.e., the ones which are not supported by the presence or absence of an AKI diagnosis on the discharge summary of a patient. *Finally*, the urine KDIGO AKI criteria was used in addition to creatinine one, which makes the labels more accurate and different from other studies that used only the latter [40][23][4][9][42].

One of the limitations of this work is the presence of noisy labels in subtasks 2 and 3, which may contribute to the models' inability to learn enough to make compelling predictions. *Second*, the small final cohort dataset and insufficient data for the deep learning model to learn efficiently. *Third*, the model was trained and evaluated on patients from only one medical facility, which may cause poor generalizability. *Finally*, the deep learning models were not extensively evaluated, meaning that more parameter tweaks, data preprocess techniques, and different architectures could be explored to assess the LSTM-based approach's performance to this task comprehensively.

Future work. From the data perspective, a possible subject for future research could be experimenting with all samples for which labels were calculated. The final cohort can be enriched by adding the samples with false positive and false negative labels (the ones defined as positive by the algorithm but did not have AKI diagnoses in the discharge records and vice versa). Since the false positive labels amounted to around 8k samples, it could significantly balance the data. Another direction for future work is to define the labels as separate, not overlapping groups, i.e., labels for stage 1 are only assigned to patients who satisfy the criteria for the 1 stage and do not satisfy the criteria for stage 2, and so on. This approach would reduce the amount of noise in the labels. However, it would reduce the number of samples for each subtask as well. Finally, more experiments could be done with different observation and prediction windows, for example, exploring the model's performance using information from the first two days in ICU to predict the next two days' AKI outcome. According to Chen et al.[8], expanding the observation window to 48 hours have a positive effect on the model's performance.

From the architecture perspective, since the LSTM-based model gets a variety of information represented as raw text, the attention mechanism could be integrated into the LSTM model to improve its ability to focus on the relevant information. Another suggestion is to experiment separately with only dynamic information (such as laboratory test results, vital signs, and medications) and explore the effect of static information (such as age, race, and previous diagnoses) on the model's performance.

Chapter 6

Conclusion

Predicting Acute Kidney Injury is a crucial yet challenging task. A number of studies explored different aspects of it: predicting an AKI onset in a long-term[1], as well as a short term perspectives [42][40]; assessing the risk of the death in AKI patients [28]; AKI stage prediction [8], and etc. In this thesis, we approach this problem by predicting the next day's AKI onset from the data recorded during the first day in the ICU. Three subtasks corresponding to different granularity levels of the predictions were studied and evaluated. Since the dataset does not contain the labels for a given task, the AKI detection algorithm was developed and used to obtain the labels for supervised models' learning. Due to the low number of samples and prevalence of negative labels, three balancing techniques and their combinations were used in the experiments.

The weighted XGBoost model combined with the undersampling technique achieved the best result in predicting the AKI onset of any stage and stages 2 and 3 together on the next day in the ICU. When it comes to an LSTM-based approach applied to this task, the performance of the LSTM-based model drops compared to the XGBoost model. However, the analysis of the LSTM-based model's explanations showed the ability to focus on the relevant information for predicting AKI.

The early prediction of AKI could be considerable support for clinicians since this would enable them to pay more attention to at-risk patients and prevent adverse outcomes. Currently, most of the AKI predicting approaches represent the EHR data as a vector of manually selected features, hence relying on the expertise of the designer of an algorithm. Applying techniques used in NLP to unstructured EHR data could help to use unstructured medical records with the least manual data manipulations.

References

[1] Sheikh S. Abdullah, Neda Rostamzadeh, Kamran Sedig, Amit X. Garg, and Eric McArthur. Predicting Acute Kidney Injury: A Machine Learning Approach Using Electronic Health Records. *Information*, 11(8):386, August 2020. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Jose Roberto Ayala Solares, Francesca Elisa Diletta Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir H. Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, Nathalie Conrad, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, 101:103337, January 2020.

[3] Samira Bell, Matthew T James, Chris K T Farmer, Zhi Tan, Nicosha de Souza, and Miles D Witham. Development and external validation of an acute kidney injury risk score for use in the general population. *Clinical Kidney Journal*, 13(3):402–412, June 2020.

[4] Samira Bell, Matthew T James, Chris K T Farmer, Zhi Tan, Nicosha de Souza, and Miles D Witham. Development and external validation of an acute kidney injury risk score for use in the general population. *Clinical Kidney Journal*, 13(3):402–412, June 2020.

[5] Rinaldo Bellomo, John A Kellum, and Claudio Ronco. Acute kidney injury. *The Lancet*, 380(9843):756–766, August 2012.

[6] Rinaldo Bellomo, Claudio Ronco, John A. Kellum, Ravindra L. Mehta, Paul Palevsky, and Acute Dialysis Quality Initiative workgroup. 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. *Critical Care (London, England)*, 8(4):R204–212, August 2004.

[7] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[8] Zhimeng Chen, Ming Chen, Xuri Sun, Xieli Guo, Qiuna Li, Yinqiong Huang, Yuren Zhang, Lianwei Wu, Yu Liu, Jinting Xu, Yuming Fang, and Xiahong Lin. Analysis of

the Impact of Medical Features and Risk Prediction of Acute Kidney Injury for Critical Patients Using Temporal Electronic Health Record Data With Attention-Based Neural Network. *Frontiers in Medicine*, 8:658665, June 2021.

[9] Peng Cheng, Lemuel R. Waitman, Yong Hu, and Mei Liu. Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate? *AMIA Annual Symposium Proceedings*, 2017:565–574, April 2018.

[10] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] John Danziger, Ken Chen, Joon Lee, Mengling Feng, Roger G. Mark, Leo Anthony Celi, and Kenneth J. Mukamal. Obesity, Acute Kidney Injury, and Mortality in Critical Illness. *Critical care medicine*, 44(2):328–334, February 2016.

[12] Joseph F. Dasta and Sandra Kane-Gill. Review of the Literature on the Costs Associated With Acute Kidney Injury. *Journal of Pharmacy Practice*, 32(3):292–302, June 2019.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[14] Martín-del-Campo Fabiola, Ruvalcaba-Contreras Neri, Velázquez-Vidaurri Alma L, Cueto-Manzano Alfonso M, Rojas-Campos Enrique, Cortés-Sanabria Laura, Espinel-Bermúdez María C, Hernández-González Sandra O, Nava-Zavala Arnulfo H, Fuentes-Orozco Clotilde, Balderas-Peña Luz, González-Ojeda Alejandro, and Mireles-Ramírez Mario. Morbid Obesity Is Associated With Mortality And Acute Kidney Injury In Hospitalized Patients With Covid-19. *Clinical Nutrition ESPEN*, September 2021.

[15] Pehuén Fernández, Emanuel J Saad, Augusto Douthat Barrionuevo, Federico A Marucco, María Celeste Heredia, Ayelén Tarditi Barra, Silvina T Rodriguez Bonazzi, Melani Zlotogora, María Antonella Correa Barovero, Sofía M Villada, Juan Pablo Maldonado, María Luján Alaye, Juan Pablo Caeiro, and Ricardo A Albertini. THE INCIDENCE, RISK FACTORS AND IMPACT OF ACUTE KIDNEY INJURY IN HOSPITALIZED PATIENTS DUE TO COVID-19. page 9.

[16] C. J. Girman, T. D. Kou, K. Brodovicz, C. M. Alexander, E. A. O'Neill, S. Engel, D. E. Williams-Herman, and L. Katz. Risk of acute renal failure in patients with Type 2 diabetes mellitus. *Diabetic Medicine: A Journal of the British Diabetic Association*, 29(5):614–621, May 2012.

[17] Seung Seok Han, Seon Ha Baek, Shin Young Ahn, Ho Jun Chin, Ki Young Na, Dong-Wan Chae, and Sejoong Kim. Anemia Is a Risk Factor for Acute Kidney Injury and Long-Term Mortality in Critically Ill Patients. *The Tohoku Journal of Experimental Medicine*, 237(4):287–295, December 2015.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] A. Hoerbst and E. Ammenwerth. Electronic Health Records. *Methods of Information in Medicine*, 49(4):320–336, 2010. Publisher: Schattauer GmbH.

[20] Institute of Medicine (US) Committee on Improving the Patient Record. *The Computer-Based Patient Record: Revised Edition: An Essential Technology for Health Care*. National Academies Press (US), Washington (DC), 1997.

[21] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. Version Number: 1.0 Type: dataset.

[22] Arif Khwaja. KDIGO clinical practice guidelines for acute kidney injury. *Nephron. Clinical Practice*, 120(4):c179–184, 2012.

[23] Kipyo Kim, Hyeonsik Yang, Jinyeong Yi, Hyung-Eun Son, Ji-Young Ryu, Yong Chul Kim, Jong Cheol Jeong, Ho Jun Chin, Ki Young Na, Dong-Wan Chae, Seung Seok Han, and Sejoong Kim. Real-Time Clinical Decision Support Based on Recurrent Neural Networks for In-Hospital Acute Kidney Injury: External Validation and Model Interpretation. *Journal of Medical Internet Research*, 23(4):e24120, April 2021. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

[25] John Lafferty, Andrew McCallum, and Fernando C N Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. page 10.

[26] Norbert H. Lameire, Arvind Bagga, Dinna Cruz, Jan De Maeseneer, Zoltan Endre, John A. Kellum, Kathleen D. Liu, Ravindra L. Mehta, Neesh Pannu, Wim Van Biesen, and Raymond Vanholder. Acute kidney injury: an increasing global concern. *The Lancet*, 382(9887):170–179, 2013.

[27] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):1–12, April 2020. Number: 1 Publisher: Nature Publishing Group.

[28] Ke Lin, Yonghua Hu, and Guilan Kong. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *International Journal of Medical Informatics*, 125:55–61, May 2019.

[29] Nuttha Lumlertgul, Monpraween Amprai, Sasipha Tachaboon, Janejira Dinhuzen, Sadudee Peerapornratana, Stephen J. Kerr, and Nattachai Srisawat. Urine Neutrophil Gelatinase-associated Lipocalin (NGAL) for Prediction of Persistent

AKI and Major Adverse Kidney Events. *Scientific Reports*, 10(1):8718, May 2020. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biomarkers;Diseases;Medical research;Nephrology Subject_term_id: biomarkers;diseases;medical-research;nephrology.

[30] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*, November 2017. arXiv: 1705.07874.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[32] D. Patschan and G. A. Müller. Acute Kidney Injury in Diabetes Mellitus. *International Journal of Nephrology*, 2016:6232909, 2016.

[33] John W. Pickering and Zoltan H. Endre. GFR shot by RIFLE: errors in staging acute kidney injury. *Lancet (London, England)*, 373(9672):1318–1319, April 2009.

[34] Nina Rank, Boris Pfahringer, Jörg Kempfert, Christof Stamm, Titus Kühne, Felix Schoenrath, Volkmar Falk, Carsten Eickhoff, and Alexander Meyer. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *npj Digital Medicine*, 3(1):1–12, October 2020. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnosis;Preventive medicine Subject_term_id: diagnosis;preventive-medicine.

[35] Adwait Ratnaparkhi. Maximum Entropy Models for Natural Language Processing. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 800–805. Springer US, Boston, MA, 2017.

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[37] Samuel A. Silver and Glenn M. Chertow. The Economic Consequences of Acute Kidney Injury. *Nephron*, 137(4):297–301, 2017.

[38] Scott M. Sutherland, Lakhmir S. Chawla, Sandra L. Kane-Gill, Raymond K. Hsu, Andrew A. Kramer, Stuart L. Goldstein, John A. Kellum, Claudio Ronco, Sean M. Bagshaw, and 15 ADQI Consensus Group. Utilizing electronic health records to predict acute kidney injury risk and outcomes: workgroup statements from the 15(th)

ADQI Consensus Conference. *Canadian Journal of Kidney Health and Disease*, 3:11, 2016.

[39] Xiaojing Tang and John C. Lieske. Acute and chronic kidney injury in nephrolithiasis. *Current opinion in nephrology and hypertension*, 23(4):385–390, July 2014.

[40] Nenad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O. Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R. Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R. Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, August 2019. Number: 7767 Publisher: Nature Publishing Group.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. arXiv: 1706.03762.

[42] Yuan Wang, Yake Wei, Hao Yang, Jingwei Li, Yubo Zhou, and Qin Wu. Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. *BMC Medical Informatics and Decision Making*, 20(1):238, September 2020.

APPENDICES

Appendix A

Python Implementation

A.1 Metrics

$$\textit{True Positive Rate} = \textit{Recall} = \textit{Sensitivity} = \frac{TP}{TP + FN}$$

$$\textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{False Positive Rate} = \frac{FP}{FP + TN}$$

$$F1\text{-}\textit{score} = \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

$$PR\ AUC = auc(\textit{Recall},\ \textit{Precision})$$

$$ROC\ AUC = auc(\textit{False Positive Rate},\ \textit{True Positive Rate})$$

where *auc(X, Y)* is a function calculating area under the curve obtained by plotting X on x-axis and Y in y-axis.

33% Precision Sensitivity is sensitivity obtained by setting the operating threshold to the value corresponding to the Precision of 33%.

A.2 Libraries

- torch==1.11.0
- xgboost==1.5.2

- matplotlib==3.5.1
- seaborn==0.11.2
- numpy==1.21.6
- pandas==1.3.5
- tokenizers==0.12.1
- lime==0.2.0.1
- sklearn==1.0.2
- imblearn==0.9.0
- wandb==0.12.11
- pickle5==0.0.12
- re==2.2.1
- tqdm==4.64.0
- pip==22.1.2

A.3 Code

The code is available and located in the following repository: `https://github.com/maslenkovas/aki_prediction.git`