Multi-Stage Chronic Kidney Disease Classification on Longitudinal Data

Ali Guran[1], Gary K.L. Tam[1], James Chess[2], and Xianghua Xie[1]

[1] Department of Computer Science, Swansea University, Swansea, UK SA1 8EN
[2] Wales Kidney Research Unit and Morriston Hospital, Swansea, UK SA6 6NL

Abstract. Chronic Kidney Disease (CKD) presents a significant global health challenge, often going unnoticed in patients until reaching advanced stages. Late-stage CKD profoundly impacts patients' lifestyles. It often necessitates weekly dialysis or kidney transplants, which both require costly medical support. Detecting early-stage CKD, however, facilitates preventive measures through lifestyle changes and medical interventions. This highlights the importance of early detection and accurate staging. Recent advancements in machine learning offer immense promise for diagnosing and identifying the CKD stages. However, most studies focus on only binary classification (CKD or not CKD) using cross-sectional data. Nonetheless, it is often observed that longitudinal analysis is more suitable for long-term disease prediction, leveraging extensive temporal data. In this study, we conducted an analysis using a comprehensive dataset of blood test results obtained from the Welsh Results Reports Service (WRRS) accessed through the Secure Anonymised Information Linkage (SAIL) Databank. By utilizing longitudinal dataset and employing machine learning techniques, namely Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM) algorithms, we present the first study to classify all five stages (from 1st to 5th) of CKD, including both early and late stages. These techniques enabled us to determine the stages of CKD in patients with precision. We also compare our models against cross-sectional techniques commonly used in the literature, namely RF, SVM, Decision Tree and Logistic Regression. Our findings indicate that the longitudinal model yields better results. This could potentially be valuable for General Practitioners (GPs) in identifying CKD early for referral.

Keywords: Chronic Kidney Disease · Machine Learning.

1 Introduction

Chronic Kidney Disease (CKD) stands as one of the most prevalent diseases today. In 2017, an analysis by the Global Burden of Disease (GBD) showed that approximately 697.5 million people worldwide were affected by CKD, including all stages of the disease [29]. In 2019, 1.43 million people died from CKD and between 1990 and 2019, CKD rose from 19th to 11th among leading causes of death [29]. Kidney Research UK reports that in 2023, the estimated annual cost to the

Table 1. Comparison of the existing works. LT: Lab Tests, PD: Previous Diagnoses of Other Conditions, CS: Cross-Sectional Dataset, L: Longitudinal Dataset, PP: Progression Prediction, BC: Binary Classification, MC: Multi-class Classification

Papers	Dataset	Size	LT	PD	CS	L	PP	BC	MC
[1, 3, 5, 6, 8, 14–16, 18, 27]	UCI	400	✓	✓	✓	-	-	✓	-
[19]	UCI	400	✓	✓	✓	-	-	-	✓
[12]	UCI	400	✓	✓	✓	-	-	✓	✓
[31]	CMH	40,000	-	✓	✓	-	-	✓	-
[23]	TNH	90,000	-	✓	✓	-	-	✓	-
[11]	GHG	400	✓	-	✓	-	-	✓	-
[4]	SPH	1718	✓	✓	✓	-	-	-	✓
[33]	EHRs	82,000	✓	-	-	✓	✓	-	-
[2]	A-CKD	3,729	✓	-	-	✓	✓	-	-
Ours	**SAIL**	**≈3000**	✓	**-**	✓	✓	**-**	**-**	✓

National Health Service (NHS) for each patient with kidney disease in the UK was £34K, with the economic burden of kidney disease in the UK reaching £7.0 billion [30]. CKD is fundamentally categorized into five different stages, ranging from 1 to 5. Late-stage (stage 4-5) CKD profoundly impacts patients' lifestyles, necessitating weekly dialysis or kidney transplant, which requires costly medical support. Detecting early-stage CKD, however, would well-equip patients and facilitate preventive measures through lifestyle changes and proactive medical interventions. The determination of the patient's stage is critical in deciding on their treatment. All these underscore the importance of early detection and accurate staging.

Applications of Machine Learning (ML) algorithms in the field of clinical medicine today yield promising results. These algorithms assist doctors in making clinical decisions regarding the detection, progression, and treatment methods of diseases. In the area of CKD, existing works can be divided into two categories: those that make predictions on cross-sectional data [1, 3–6, 8, 11, 12, 14–16, 18, 19, 23, 27, 31] and those on longitudinal data [2, 28, 33]. In cross-sectional work, many works focus solely on binary classification (CKD or not CKD) [1, 3, 5, 6, 8, 11, 12, 14–16, 18, 23, 27, 31], very few undertake multi-stage classification of all stages of CKD [4, 12, 19]. In longitudinal studies, the primary focus has been on predicting the progression of chronic kidney disease (CKD). Recent research [33] specifically addresses predicting the transition from stage 2/3 to 4/5 of the disease. Similarly, [2] predicts the necessity of referring CKD patients from primary care to secondary care. Earlier studies, discussed in the review [28], include various aspects such as predicting the onset of CKD and forecasting the decline in kidney function. The majority of other studies revolve around predicting the progression to end-stage kidney disease (ESKD). To our understanding, there is no CKD longitudinal prediction technique that leverages longitudinal data that classifies all stages of CKD (Table 1). It is however commonly recognized that longitudinal analysis is better suited for long-term disease prediction and allows for better predictions, utilizing extensive temporal data [32].

The scarcity of quality longitudinal CKD datasets significantly hampers progress in this field. Many publicly available datasets are limited to cross-sectional data only. Some encompass demographic information, current diagnoses, and lab tests. Others focus solely on demographic data and blood tests (Table 1). However, relying solely on cross-sectional data, particularly single-day blood tests, is insufficient for accurate CKD staging. According to Kidney Disease: Improving Global Outcomes (KDIGO) guidelines [29], CKD diagnosis requires persistent kidney abnormalities over at least 90 days. This underscores the necessity of longitudinal data analysis for reliable predictions. Longitudinal analysis, based on parameters like estimated glomerular filtration rate (eGFR) over time, provides a more robust approach to CKD staging compared to cross-sectional data. Furthermore, the dataset limitation extends to the scarcity of multi-stage datasets. Few studies undertake multiclass classification, often resorting to converting binary class labels into multiclass categories based on eGFR values [12, 19]. However, this approach that relies on single observations rather than multiple observations spanning 90 days may not follow KDIGO guidelines [29]. To our knowledge, there is a lack of proper longitudinal CKD dataset that supports multistage classification of all 5 stages.

In this study, we aimed to bridge this gap by conducting multi-stage classification predictions on longitudinal data. Firstly, we obtained access to the SAIL databank (Section 3), which houses anonymized Welsh blood sample data. Following the guidelines provided by the KDIGO CKD Work Group [29], we labeled all 5 CKD stages based on patients' eGFR observations over the last 90 days. Secondly, we developed machine learning techniques, namely LSTM and Bidirectional LSTM, to effectively handle such multivariate longitudinal dataset.

We summarize our main contributions below:

- We introduce the first study on predicting multi-stage (on all five stages) of CKD through developing LSTM and Bidirectional-LSTM models trained on multivariate longitudinal data for accurate CKD stage prediction.
- We curate our dataset carefully in accordance with the guidelines provided by KDIGO [29], utilizing observations from the last 90 days to ensure reliability.
- We also present models for making cross-sectional predictions, enabling a fair comparison with our longitudinal models.
- Our results demonstrate that longitudinal techniques, which account for disease dynamics, yield better performance in predicting stages of CKD.

2 Related Works

Various Machine Learning methods have been employed for classifying Chronic Kidney Disease stages. The majority of studies focus on binary classification (CKD vs non-CKD) using cross-sectional data. Only a few attempt multiclass classification (all 1-5 stages of CKD) predictions on cross-sectional data. However, to date, no study has performed multiclass classification using longitudinal patient data. A summary of these studies is provided in Table 1. We discuss these methods and datasets below.

2.1 Methods

Binary Classification Methods Salekin et al. [27] developed a Machine Learning classifier using k-nearest neighbors, random forests, and neural networks for the detection of Chronic Kidney Disease (CKD vs. non-CKD). The random forest algorithm achieved an accuracy of 99.3% in this binary classification task. In 2018, Aljaaf et al. conducted binary classifications through various machine-learning algorithms for CKD detection, including classification and regression tree (RPART), support vector machine (SVM), logistic regression (LOGR), and multilayer perceptron neural network (MLP) methods [1]. In 2020, Khan et al. classified patients into CKD and non-CKD using seven different models, including Support Vector Machine, NBTree, Logistic Regression, Naïve Bayes, Multi-layer Perceptron, J48, and Composite Hypercube on Iterated Random Projection (CHIRP) [15]. The accuracies of these models ranged from 95% to 99.75%, varying among these studies. While these works perform well in classifying CKD vs non-CKD, they do not consider multi-stage (5) classification.

Multi-Stage (5 Classes) Classification Methods In 2019, Rady et al. employed Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Radial Basis Function (RBF) algorithms to perform multi-stage classifications on cross-sectional data [19]. In 2021, Ilyas et al. utilized the Random Forest and J48 (C4.5) algorithms to classify patients into various CKD stages [12]. The above work both used the UCI dataset, with accuracy ranging from 51.5% to 96%, depending on the models used. In 2022, Debal et al. conducted both binary classification and multiclass classifications using Random Forest, Support Vector Machine, and Decision Tree algorithms [4]. The highest accuracies for multiclass classification were 79.0% for Random Forest, 63% for Support Vector Machine, and 78% for Decision Tree on SPH dataset. While these attempts aim to predict multi-stage classification, they are mostly based on cross-sectional data only. There is no study that considers multi-stage classification using longitudinal data, which our study aims to address.

Longitudinal Methods Some studies have explored machine learning techniques applied to longitudinal data. Earlier studies, as discussed in the review [28], employed longitudinal methods to make predictions regarding progression. These predictions include forecasting whether patients will develop CKD in the future, estimating the loss of kidney function, and predicting progression to end-stage disease, rather than classifying CKD stages (1 to 5). In 2020, Au-Yeung et al. [2] aimed to assist kidney teams in advising primary care GPs to refer patients to secondary care earlier for specialist assessment and medical intervention. They employed longitudinal eGFR readings to classify whether patients should be referred to secondary care, constituting a binary classification task. Another recent study by Zhu et al. [33] focused on predicting CKD progression from stages 2/3 to stages 4/5. Patients progressing to stages 4/5 within a specified window are labeled as 1, while those not progressing are labeled 0, resulting in binary classification. Both studies primarily address binary classification tasks on longitudinal data. Notably, neither study categorizes patients according to their specific CKD stages (1 to 5). To our knowledge, there is no study examining the classification of all CKD stages using a multivariate longitudinal dataset. This study addresses this research gap.

2.2 Existing Datasets

Many existing cross-sectional datasets are designed for binary classification. The UCI dataset [22], commonly utilized in numerous existing studies, is provided by the University of California Irvine (UCI). It comprises 400 instances and 25 attributes including age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia. This cross-sectional dataset contains only two classes: CKD and non-CKD. The SPH dataset utilized in the study by Debal et al. [4] was collected between 2018 and 2019 at St. Paulo's Hospital. It comprises 1718 instances and 19 variables, including age, gender, blood pressure, specific gravity, chloride, sodium, potassium, blood urea nitrogen, serum creatinine, hemoglobin, red blood cell count, white blood cell count, mean cell volume, platelet count, hypertension, diabetes mellitus, anemia, and heart disease. While it is a dataset with multiclass labels, it is not longitudinal. Iliyas et al. used another cross-sectional GHG dataset in their study [11], created by the General Hospital in Gashua Local Government Area of Yobe State. This dataset comprises 400 instances with 11 variables, similar to those in other datasets, and includes a target variable classified as CKD and non-CKD. Two additional datasets [23, 31] contain only previous diagnoses of patients rather than blood tests and utilize binary labels. The primary issues with these cross-sectional datasets are their unsuitability for longitudinal analysis, reliance on single-day tests for patient labeling, and the presence of imbalanced data. Earlier studies using longitudinal datasets [28] have focused on predicting CKD progression rather than staging the disease (from 1 to 5). These datasets vary in size, from 465 to 14,039 subjects, and may include diverse demographic details, vital signs, laboratory results, and health behaviors. Data are usually sourced from medical centers or renal units, requiring approval from respective authorities (e.g., US, Japan). However, without direct access to the data, it remains uncertain whether these datasets can be effectively utilized for predicting all five stages.

3 Dataset

Our data was provided by the Secure Anonymised Information Linkage (SAIL) Databank [7, 13, 17, 20, 21], a secure research environment in Wales funded by Health and Care Research Wales, based at Swansea University's Medical School. SAIL facilitates remote access, linkage, and analysis of administrative and health data, fostering collaborations with research groups across Wales and the UK. The

Table 2. CKD Stages [29].

Stage	Description	eGFR Range
1	Possible kidney damage with normal kidney function	≥ 90
2	Kidney damage with mild loss of kidney function	60-89
3	Mild to severe loss of kidney function	30-59
4	Severe loss of kidney function	15-29
5	Kidney failure	< 15

datasets in SAIL are anonymized and linked using a split file process, ensuring that identifiable data and clinical information are not accessible simultaneously.

We utilize the Welsh Results Reports Service (WRRS) and Welsh Longitudinal General Practice Dataset (WLGP) - Welsh Primary Care datasets provided by SAIL. The WRRS dataset encompasses observation requests, results, and reports, representing a population size of over 4 million. Healthcare professionals (HCPs) throughout Wales have the capability to access, input, and review laboratory results, including blood tests, for pathology requests and related outcomes across primary and secondary care settings. This accessibility is facilitated by the Welsh Results Reporting Service (WRRS), providing a dedicated platform for patients to undergo blood testing at mobile units or local centers, while enabling clinicians to promptly access the results remotely. The service is designed to achieve time savings, reduce test duplication, and enhance patient safety [25]. The WLGP includes 80% of Wales' GP practices and 83% of the country's total population. It is linkable to other datasets, such as custom project-specific cohorts, using anonymized variables for patients and general practitioners. A clinical information system is used by every GP office to keep an electronic health record for every patient. This record includes all signs, symptoms, test results, diagnoses, prescription treatments, referrals for specialized care, and social factors related to the patient's home environment representing a population size of $\approx$3.5 million. During the patient consultation, the clinician enters most of the data [24]. We link these two datasets using anonymized patient variables. We obtain the blood test from the WRRS and the patient's demographic information such as age and gender from the WLGP respectively.

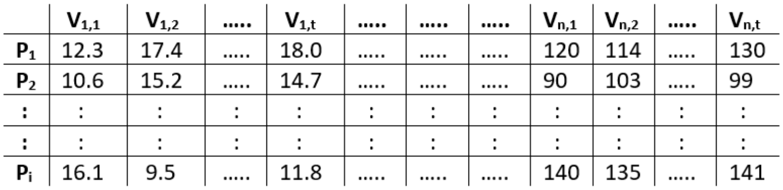
3.1 Data Processing

In our study, we curated both cross-sectional and longitudinal models for comparison. As the requirements for each differ, we created two separate datasets. The first dataset is longitudinal, while the other is cross-sectional. To train both types of models, we included the following attributes: *patients' age*, *gender*, and various blood tests containing *creatinine*, *albumin*, *red blood cell count*, *white blood cell count*, *alkaline phosphatase*, *alanine transaminase, sodium, potassium, mean cell haemoglobin*, *mean cell volume*, *total protein*, and *globulin*. These attributes and the setup bear similarity to those employed in established studies. (Section 2.2).

Longitudinal Dataset To construct our longitudinal dataset, we retrieved all records containing the specified blood tests for each patient to avoid missing values, ensuring data completeness. Patients missing any test were excluded from the dataset. Consequently, we obtained a dataset devoid of missing values, utilizing only actual records, thereby bolstering dataset reliability. Longitudinal data were collected from all dates with complete test records, providing longitudinal data for all patients. However, a challenge arises from the varying testing frequency and duration among patients, which is common due to differences in healthcare needs and schedules. Statistics showed a mean of 2460 days of records per patient, prompting us to filter for patients with at least 2460 days of data.

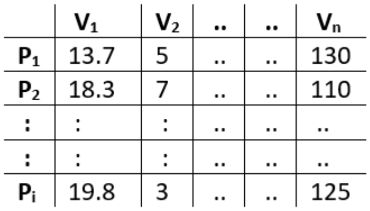
To address variations in the number of tests on different dates for each patient, we imputed missing test values using linear interpolation for each day. Then, all longitudinal data are aligned to the last reading, following Au-Yeung et al.'s method [2]. Subsequently, we retained imputed tests for each patient at 10-day intervals over the last 2460 days, resulting in 247 observations for each test. Patients were categorized into 5 CKD stages based on their eGFR values over the last 90 days [29]. The eGFR value ranges for each CKD Stage are provided in Table 2. Those with eGFR values outside the CKD stage range were excluded. When all obtained patients were labelled, it was observed that the CKD stage with the lowest number of patients had ≈600 patients (the exact numbers of patients are not disclosed here in compliance with SAIL output review policy [26]). Taking into account the data distribution across stages, we opted to randomly choose ≈600 patients from each stage. This approach was necessary as other stages had a higher patient count, which could have skewed the balance of data across stages. Hence, we acquired a total of ≈3000 patients across all stages of chronic kidney disease (CKD). An illustrative example of the multivariate longitudinal data structure is shown in Figure 1. In the table, P_i represents each patient, and $V_{n,t}$ represents variables within the multivariate longitudinal dataset, where n refers to each attribute and t denotes observations in the time domain. This multivariate longitudinal dataset enables us to evaluate the effectiveness of our longitudinal techniques and implementations for predicting multi-stage CKD.

Cross-Sectional Dataset To create the cross-sectional dataset, we included the most recent cross-sectional (time point) blood tests of each patient from our longitudinal dataset and the respective CKD stage of each patient. In other words, we included a total of ≈3000 patients and their blood tests, ≈600 patients from each stage, in our cross-sectional dataset. Our purpose in doing this is



	$V_{1,1}$	$V_{1,2}$	.....	$V_{1,t}$	.....	.....	.....	$V_{n,1}$	$V_{n,2}$	.....	$V_{n,t}$
P_1	12.3	17.4	.....	18.0	.....	.....	.....	120	114	.....	130
P_2	10.6	15.2	.....	14.7	.....	.....	.....	90	103	.....	99
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:
P_i	16.1	9.5	.....	11.8	.....	.....	.....	140	135	.....	141

Fig. 1. Representative Longitudinal Data Structure (unreal values for illustration)



	V_1	V_2	..	..	V_n
P_1	13.7	5	..	..	130
P_2	18.3	7	..	..	110
:	:	:	..	..	..
:	:	:	..	..	..
P_i	19.8	3	..	..	125

Fig. 2. Representative Cross-Sectional Data Structure (unreal values for illustration)

to compare our longitudinal models with our cross-sectional models fairly. An illustrative example of the multivariate cross-sectional data structure we acquired is depicted in Figure 2. In the table, P_i represents each patient, and V_n represents each attribute, such as creatinine and albumin.

4 Methods

We implement and compare common machine learning techniques utilized in CKD literature on our cross-sectional dataset. These include Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) and Logistic Regression (LR) algorithms. The grid search method is employed for hyperparameter tuning for the models. Further, we implement two longitudinal models, namely Long Short-Term Memory (LSTM) and Bi-directional Long Short-Term Memory (Bi-LSTM) algorithms, and evaluate them on our longitudinal CKD dataset.

4.1 Cross-Sectional Models

Random Forest (RF) is an ensemble learning method. It excels in classification tasks and feature selection, exhibiting lower overfitting compared to other algorithms. By forming a "forest" of decision trees and averaging them, it enhances prediction accuracy, especially for complex datasets [10]. It is adept at handling challenges like multiple class categorization. Its capability to manage thousands of input variables is invaluable for analyzing cross-sectional data.

Support Vector Machine (SVM) is particularly effective for classification tasks in high-dimensional spaces, making it well-suited for multivariate cross-sectional data with numerous variables. Its ability to identify optimal hyperplanes for class division allows it to handle complex decision boundaries, crucial for classification. Its robustness and generalization to new data stem from its focus on maximizing the margin between classes, ensuring reliable performance on both new and unseen data [10].

Decision Trees (DT) are ideal for multiclass classification on multivariate cross-sectional data due to their intuitive design, facilitating easy understanding and visualization of decision-making processes. Their flexibility in handling both numerical and categorical data makes them suitable for multivariate datasets.

The hierarchical structure of decision trees effectively captures the complexity of multiclass problems, offering clear insights into the significance of various variables for classification. Despite their tendency to overfit, techniques such as pruning and setting maximum depth can mitigate this issue, making Decision Trees a robust and interpretable tool for assessing cross-sectional data in a multiclass environment [10].

Logistic Regression (LR) offers a simple, easily interpretable model, particularly advantageous in multiclass classification tasks where understanding predictor effects is crucial. Its coefficients can be directly interpreted as odds ratios, providing clarity on variable influences [10]. Compared to more complex models, logistic regression models are less prone to overfitting due to their lower variance. Regularization techniques like L1, L2, or a combination can be applied to further prevent overfitting and enhance performance on unseen data, especially important in multivariate contexts with numerous predictors.

4.2 Longitudinal Models

Long Short-Term Memory (LSTM) excels in multiclass classification tasks on multivariate longitudinal data due to their specialized architecture tailored for handling sequential data across time. They effectively capture temporal relationships and patterns, crucial for understanding the evolution of multiclass states over numerous time points. LSTMs integrate prior data points into current predictions, leveraging their long-term memory and information processing capabilities. Their sophisticated understanding of sequential data enables them to account for changes and trends within variables over time, making them effective in utilizing rich time-series data in longitudinal datasets [9]. CKD progression typically occurs gradually, emphasizing the importance of long-term health data analysis for early disease detection. By effectively capturing critical signals and temporal variations, LSTM models can offer accurate predictions even during the early stages of CKD. We hypothesize that LSTM will outperform simpler cross-sectional techniques in predicting early CKD stages.

A multivariate time series classification LSTM model was built using Keras. It includes two LSTM layers with 16 and 8 hidden units respectively, each followed by a 20% Dropout layer. The output layer consists of 5 neurons with softmax activation for predicted class probabilities. In our implementation, the model was trained for up to 50 epochs, and batches of 8 samples were used in each training cycle.

Bi-directional Long Short-Term Memory (Bi-LSTM) significantly enhances multiclass classification performance on multivariate longitudinal data by leveraging both forward and backward sequential contexts. They capture emerging patterns and relationships over time from past and future contexts, providing a comprehensive understanding of temporal dynamics. Integrating data across the full sequence length, Bi-LSTMs excel in complex classification scenarios, considering the complete sequence for each prediction. They offer a useful framework for modeling temporal interactions and dependencies between variables in both directions, enabling deeper understanding and precise predictions in multiclass classification tasks [9].

We utilized Bi-LSTM due to its capability to analyze sequential data in both forward and backward directions. This approach is crucial for minimizing misclassifications between adjacent disease stages. By capturing intricate patterns in CKD progression across various stages, Bi-LSTM enables a more detailed understanding of disease dynamics. Our hypothesis is that Bi-LSTM's bidirectional configuration will reduce misclassifications between adjacent CKD stages compared to unidirectional LSTM. This improvement may enhance disease management by facilitating more accurate diagnosis and intervention.

A classification model for multivariate longitudinal data was constructed using Bi-LSTM networks in Keras. It features two bidirectional LSTM layers with 16 and 8 hidden units respectively, along with Dropout (20%) to mitigate overfitting. The model includes a Dense output layer with softmax activation. Compiled with Adam optimizer, it was trained for 50 epochs with a batch size of 8 samples.

5 Experiments and Results

5.1 Experiment Setup

To ensure accurate model evaluation, we employed k-fold cross-validation. This technique assesses how well the model fits the data and detects issues like overfitting. In k-fold cross-validation, the dataset is divided into subsets (folds), with each subset used alternately as the test set while the remaining parts serve as the training set. This process is repeated for k folds, aiming to provide reliable generalization performance estimates across different data subsets. In our study, we utilized (k=5) 5-fold cross-validation, a widely preferred balance between computational cost and reliability. In each fold, the data is split into 80% training and 20% testing sets to ensure ample training data for model learning while maintaining an objective evaluation of its performance on independent data.

5.2 Evaluation Metrics

We employed the following various evaluation metrics, where TP: Correctly predicted positive, TN: Correctly predicted negative, FP: Incorrectly predicted positive, FN: Incorrectly predicted negative.
Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$ indicates the proportion of all correct predictions made by the model.
Precision $= \frac{TP}{TP+FP}$ measures the proportion of positive cases predicted correctly among all cases predicted as positive.
Recall (aka. Sensitivity) $= \frac{TP}{TP+FN}$ represents the proportion of true positive cases predicted correctly among all actual positive cases.
F-Measure (aka. F1-score) $= 2 * \frac{Precision*Recall}{Precision+Recall}$ is the harmonic mean of precision and recall. It provides a balanced measure of both metrics.
Specificity $= \frac{TN}{TN+FP}$ measures the proportion of true negative cases predicted correctly among all actual negative cases.

Table 3. Multi-stage Classification Results

Cross-Sectional	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Specificity**
RF	92.74	92.72	92.74	92.73	98.18
SVM	94.93	94.94	94.94	94.94	98.73
DT	95.95	95.97	95.95	95.96	98.99
LR	96.28	96.31	96.29	96.30	99.07
Longitudinal	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Specificity**
LSTM	96.62	96.67	96.63	96.65	99.16
Bi-LSTM	**97.30**	**97.31**	**97.30**	**97.30**	**99.32**

5.3 Results

Overall Performance The overall results are presented in Table 3. Higher values of these metrics indicate better performance of the techniques. Generally, longitudinal techniques outperform cross-sectional techniques. Specifically, Bi-directional LSTM demonstrates better performance compared to LSTM.

Confusion matrix We generate a 5-by-5 confusion matrix for each technique, with each row and column representing the classification performance for the respective stages (1-5). However, adhering to SAIL Output Review Policy [26], we are unable to present the confusion matrices due to the risk of disclosing sensitive patient information, as some cells contain values equal to or less than 5. Nevertheless, to ensure a thorough analysis, we provide a summary of our observations below.

When comparing models trained on cross-sectional data (DT, LR, SVM, RF), we observe consistent performance across most stages. Generally, they predict Stage 1 and Stage 5 with higher accuracy, indicating potentially distinct characteristics in these stages. In the confusion matrix of RF, incorrect predictions are confined to the subdiagonal and superdiagonal elements. RF often makes more errors in Stage 2 and Stage 3 classifications by incorrectly categorizing instances as the nearest neighbor classes. For example, Stage 2 instances may be mistakenly labeled as either Stage 1 or Stage 3 by the model. In contrast, DT and LR demonstrate balanced errors and stable performance across stages, leading to higher overall accuracy in Table 3.

When comparing LSTM and Bi-LSTM models trained on longitudinal data, Bi-LSTM generally outperforms LSTM across all phases except for Stage 1. This highlights Bi-LSTM's superior ability to capture changes in disease stages over time, owing to its bidirectional processing capability. However, it is worth noting that LSTM performs better than Bi-LSTM in Stage 1. This may indicate that certain relevant information for predicting outcomes in Stage 1 primarily resides in the past context of the sequence, making the unidirectional LSTM architecture more effective in this particular scenario. It also suggests that different model structures may be more effective in specific stages. Similarly, incorrect predictions for LSTM and Bi-LSTM models are confined to the subdiagonal and superdiagonal elements of the confusion matrices. These misclassifications are however fewer compared to cross-sectional techniques.

Comparing cross-sectional and longitudinal models, we observe that longitudinal models outperform in Stages 1,3,4,5, likely due to their superior representation of temporal changes in the disease process. Among these, Bi-LSTM exhibits the highest accuracy, emphasizing the significance of time series data in CKD staging, and reducing misclassification between adjacent stages. For Stage 2, DT and LR are slightly better over the longitudinal models, albeit by a very narrow margin.

In conclusion, Bi-LSTM shows superior performance and consistency in CKD stage estimation, although other models also yield competitive results on specific stages. Different models perform variably based on data characteristics at various disease stages, highlighting the need for careful model selection depending on the application context and goals.

6 Conslusion

In this paper, we introduce the first study focusing on predicting all five stages of CKD by leveraging LSTM and Bidirectional-LSTM models trained on multivariate longitudinal data. Our dataset curation adheres closely to KDIGO guidelines [29], utilizing observations from the last 90 days for reliability. Furthermore, we offer models for cross-sectional predictions to ensure a fair comparison with our longitudinal counterparts. Our findings confirm our hypothesis: models trained on longitudinal data exhibit superior accuracy compared to those on cross-sectional data. This underscores the significance of capturing CKD progression over time, aligning with our initial expectations. Longitudinal techniques LSTM and Bi-LSTM prove particularly adept at capturing disease dynamics. In future work, we aim to conduct an in-depth examination and comparison of our findings on a larger dataset, employing dedicated imputation and stratification techniques. We will also explore the integration of more sophisticated deep learning models to further enhance prediction and forecasting accuracy.

Acknowledgements Ali Guran was financially supported by the Turkish Ministry of National Education, Republic of Turkiye for the Postgraduate Study Abroad Program. Gary Tam was supported by the International Mobility Award [62] and Collaboration and Knowledge Exchange Support [92S], which were funded by CHERISH-DE Centre EP/M022722/1. This project is partially supported by a fee waiver fund by SAIL. This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who make anonymised data available for research. The responsibility for the interpretation of the information we supplied is the authors' alone. All research conducted has been completed under the permission and approval of the SAIL independent Information Governance Review Panel (IGRP) project number 1220. For the purpose of Open Access the author has applied a CC BY copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Aljaaf, A.J., Al-Jumeily, D., Haglan, H.M., Alloghani, M., Baker, T., Hussain, A.J., Mustafina, J.: Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In: IEEE congress on evolutionary computation (CEC). pp. 1–9 (2018)
2. Au-Yeung, L., Xie, X., Chess, J., Scale, T.: Using machine learning to refer patients with chronic kidney disease to secondary care. In: International Conference on Pattern Recognition (ICPR). pp. 10219–10226 (2021)
3. Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., Jasiński, M., Jasiński, Ł., Gono, R., Jasińska, E., et al.: Prediction of chronic kidney disease-a machine learning perspective. IEEE Access **9**, 17312–17334 (2021)
4. Debal, D.A., Sitote, T.M.: Chronic kidney disease prediction using machine learning techniques. Journal of Big Data **9**(1), 109 (2022)
5. Devika, R., Avilala, S.V., Subramaniyaswamy, V.: Comparative study of classifier for chronic kidney disease prediction using naive bayes, knn and random forest. In: International conference on computing methodologies and communication (ICCMC). pp. 679–684 (2019)
6. Elkholy, S.M.M., Rezk, A., Saleh, A.A.E.F.: Early prediction of chronic kidney disease using deep belief network. IEEE Access **9**, 135542–135549 (2021)
7. Ford, D.V., Jones, K.H., Verplancke, J.P., Lyons, R.A., John, G., Brown, G., Brooks, C.J., Thompson, S., Bodger, O., Couch, T., et al.: The sail databank: building a national architecture for e-health research and evaluation. BMC health services research **9**, 1–12 (2009)
8. Ghosh, P., Shamrat, F.J.M., Shultana, S., Afrin, S., Anjum, A.A., Khan, A.A.: Optimization of prediction method of chronic kidney disease using machine learning algorithm. In: International joint symposium on artificial intelligence and natural language processing (iSAI-NLP). pp. 1–6 (2020)
9. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5), 602–610 (2005)
10. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction, vol. 2 (2009)
11. Iliyas, I.I., Saidu, I.R., Dauda, A.B., Tasiu, S.: Prediction of chronic kidney disease using deep neural network. arXiv preprint arXiv:2012.12089 (2020)
12. Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M.T., Iftikhar, M., Malik, M.H.: Chronic kidney disease diagnosis using decision tree algorithms. BMC nephrology **22**(1), 273 (2021)
13. Jones, K.H., Ford, D.V., Jones, C., Dsilva, R., Thompson, S., Brooks, C.J., Heaven, M.L., Thayer, D.S., McNerney, C.L., Lyons, R.A.: A case study of the secure anonymous information linkage (sail) gateway: a privacy-protecting remote access system for health-related research and evaluation. Journal of biomedical informatics **50**, 196–204 (2014)
14. Kaur, C., Kumar, M.S., Anjum, A., Binda, M., Mallu, M.R., Al Ansari, M.S.: Chronic kidney disease prediction using machine learning. Journal of Advances in Information Technology **14**(2), 384–391 (2023)
15. Khan, B., Naseem, R., Muhammad, F., Abbas, G., Kim, S.: An empirical evaluation of machine learning techniques for chronic kidney disease prophecy. IEEE Access **8**, 55012–55022 (2020)

16. Kriplani, H., Patel, B., Roy, S.: Prediction of chronic kidney diseases using deep artificial neural network technique. In: Computer aided intervention and diagnostics in clinical and medical images. pp. 179–187 (2019)
17. Lyons, R.A., Jones, K.H., John, G., Brooks, C.J., Verplancke, J.P., Ford, D.V., Brown, G., Leake, K.: The sail databank: linking multiple health and social care datasets. BMC medical informatics and decision making **9**, 1–8 (2009)
18. Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., Chen, B.: A machine learning methodology for diagnosing chronic kidney disease. IEEE Access **8**, 20991–21002 (2019)
19. Rady, E.H.A., Anwar, A.S.: Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked **15**, 100178 (2019)
20. Rodgers, S.E., Demmler, J.C., Dsilva, R., Lyons, R.A.: Protecting health data privacy while using residence-based environment and demographic data. Health & place **18**(2), 209–217 (2012)
21. Rodgers, S.E., Lyons, R.A., Dsilva, R., Jones, K.H., Brooks, C.J., Ford, D.V., John, G., Verplancke, J.P.: Residential anonymous linking fields (ralfs): a novel information infrastructure to study the interaction between the environment and individuals' health. Journal of Public Health **31**(4), 582–588 (2009)
22. Rubini L., S.P., P., E.: Chronic Kidney Disease. UCI Machine Learning Repository (2015), DOI: https://doi.org/10.24432/C5G020
23. Saif, D., Sarhan, A.M., Elshennawy, N.M.: Deep-kidney: an effective deep learning framework for chronic kidney disease prediction. Health Information Science and Systems **12**(1), 3 (2023)
24. SAIL: Welsh longitudinal general practice dataset (wlgp) - welsh primary care version 21.0.0 (2021)
25. SAIL: Welsh results reports service (wrrs) version 7.0.0 (2021)
26. SAIL Databank: Sail-pol-024 output review policy. PDF (2022), https://saildatabank.com/wp-content/uploads/2022/08/SAIL-POL-024-Output-Review-Policy-v1.2-3.pdf
27. Salekin, A., Stankovic, J.: Detection of chronic kidney disease and selecting important predictive attributes. In: IEEE International Conference on Healthcare Informatics (ICHI). pp. 262–270 (2016)
28. Schena, F.P., Anelli, V.W., Abbrescia, D.I., Di Noia, T.: Prediction of chronic kidney disease and its progression by artificial intelligence algorithms. Journal of Nephrology **35**(8), 1953–1971 (2022)
29. Stevens, P.E., Ahmed, S.B., Carrero, J.J., Foster, B., Francis, A., Hall, R.K., Herrington, W.G., Hill, G., Inker, L.A., Kazancıoğlu, R., et al.: Kdigo 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney international **105**(4), S117–S314 (2024)
30. UK, K.R.: Kidney disease: A uk public health emergency. PDF (June 2023), https://www.kidneyresearchuk.org/wp-content/uploads/2023/06/Economics-of-Kidney-Disease-full-report_accessible.pdf
31. Vásquez-Morales, G.R., Martinez-Monterrubio, S.M., Moreno-Ger, P., Recio-Garcia, J.A.: Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning. IEEE Access **7**, 152900–152910 (2019)
32. Zhao, J., Feng, Q., Wu, P., Lupu, R.A., Wilke, R.A., Wells, Q.S., Denny, J.C., Wei, W.Q.: Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Scientific reports **9**(1), 717 (2019)
33. Zhu, Y., Bi, D., Saunders, M., Ji, Y.: Prediction of chronic kidney disease progression using recurrent neural network and electronic health records. Scientific Reports **13**(1), 22091 (2023)