An LSTM-based Deep Sequential Framework for Predicting Chronic Kidney Disease Progression using Longitudinal Clinical Data

Md. Taukir Ahmed
IICT, RUET
Rajshahi-6204
Bangladesh

Mst.Suraiya Sultana
IICT, RUET
Rajshahi-6204
Bangladesh

Rubait Hasan Safiq
EEE, Varendra University
Rajshahi-6204
Bangladesh

ABSTRACT
Chronic Kidney Disease (CKD) is a progressive disorder that gradually impairs kidney function and often remains undetected until advanced stages, making early prediction essential for timely clinical intervention. Conventional machine learning methods generally struggle to model the temporal dependencies present in longitudinal patient records, resulting in limited prediction performance. This study proposes a Long Short-Term Memory (LSTM)-based deep sequential framework for predicting CKD progression using multivariate time-series clinical data. The proposed model utilizes longitudinal features, including estimated glomerular filtration rate (eGFR), serum creatinine, blood pressure, glucose, albumin, and hemoglobin, to learn complex temporal patterns associated with kidney function decline. Data preprocessing involves missing-value imputation, normalization, outlier removal, temporal feature engineering, and sliding-window sequence generation. To improve model robustness and generalization, dropout and L2 regularization are incorporated together with the Adam optimizer and Huber loss. Experimental results demonstrate that the proposed model achieves a Mean Absolute Error (MAE) of 3.08 mL/min/1.73 m², a Root Mean Square Error (RMSE) of 4.11 mL/min/1.73 m², a Mean Absolute Percentage Error (MAPE) of 6.35%, and a coefficient of determination (R^2) of 0.93. Comparative analysis with Linear Regression, Random Forest, and Support Vector Machine demonstrates the superior predictive capability of the proposed framework. These findings indicate that the proposed LSTM model provides an accurate and reliable solution for intelligent CKD progression prediction and can effectively support early diagnosis, personalized treatment planning, and clinical decision-making in modern healthcare systems.

General Terms
Time Series Analysis, Healthcare Informatics, Predictive Modeling, Deep Learning etc.

Keywords
Chronic Kidney Disease (CKD), Long Short-Term Memory (LSTM), Deep Learning, Time-Series Prediction, Disease Progression, Clinical Decision Support, eGFR (estimated glomerular filtration rate), Machine Learning

1. INTRODUCTION
Chronic Kidney Disease is a progressive and irreversible medical condition characterized by the gradual loss of kidney function over time. It has emerged as a major global public health challenge due to its high prevalence, increasing incidence, and strong association with other chronic diseases such as diabetes, hypertension, and cardiovascular disorders. According to a large-scale meta-analysis, the global prevalence of CKD is estimated to be approximately 11–13%, affecting hundreds of millions of individuals worldwide [1]. More recent findings from the Global Burden of Disease (GBD) study indicate that CKD affected approximately 788 million people globally in 2023, making it one of the leading causes of morbidity and mortality [2]. The disease is often asymptomatic in its early stages, which results in delayed diagnosis and increases the risk of progression to end-stage renal disease (ESRD), requiring dialysis or kidney transplantation [3]. CKD is typically diagnosed using biomarkers such as estimated glomerular filtration rate (eGFR) and albuminuria levels. A sustained reduction in eGFR below 60 mL/min/1.73 m² for at least three months is considered indicative of CKD [4]. However, these conventional diagnostic methods rely on static measurements and do not effectively capture the dynamic and temporal progression of the disease. Since CKD evolves gradually over time, understanding its progression requires the analysis of longitudinal patient data, including repeated clinical measurements collected over extended periods [5]. This highlights the need for advanced predictive models that can incorporate temporal dependencies and provide accurate forecasts of disease progression. In recent years, artificial intelligence (AI) and machine learning (ML) techniques have gained significant attention in the field of healthcare for disease prediction and prognosis. Traditional machine learning models such as logistic regression, decision trees, and support vector machines have been widely used for CKD classification and diagnosis [6]. Although these methods can achieve reasonable accuracy, they generally treat each data point independently and fail to account for temporal relationships present in sequential medical data. As a result, their ability to model disease progression over time is limited. Deep learning approaches, particularly recurrent neural networks (RNNs), have shown superior performance in handling time-series data due to their ability to model sequential dependencies. Among these, Long Short-Term Memory (LSTM) networks have emerged as one of the most effective architectures for sequential modeling tasks. LSTM networks address the limitations of traditional RNNs, such as the vanishing gradient problem, by introducing memory cells and gating mechanisms that enable the retention of long-term information [7]. This makes LSTM particularly suitable for analyzing longitudinal healthcare data, where past observations play a crucial role in predicting future outcomes. Several studies have demonstrated the effectiveness of LSTM models in various medical applications, including disease prediction, patient monitoring, and clinical decision support systems. For example, LSTM-based models have been successfully applied to predict

cardiovascular events, diabetes progression, and patient mortality using electronic health records (EHRs) [8]. In the context of CKD, recent research has explored the use of deep learning techniques to predict disease onset and progression. These studies highlight that incorporating temporal information significantly improves prediction accuracy compared to traditional machine learning approaches [9].Time-series modeling of CKD progression involves analyzing sequential clinical data such as eGFR, serum creatinine, blood pressure, and other biomarkers collected over time. Among these, eGFR is considered one of the most important indicators of kidney function and is widely used to monitor disease progression. A declining trend in eGFR values indicates worsening kidney function and increased risk of ESRD [10]. Therefore, accurately predicting future eGFR values can provide valuable insights into the trajectory of CKD progression and enable early intervention. Despite the growing adoption of deep learning techniques in healthcare, several challenges remain in modeling CKD progression. One of the key challenges is the availability and quality of longitudinal data. Medical datasets often contain missing values, irregular time intervals, and noise, which can affect model performance [11]. Additionally, patient heterogeneity, including differences in demographics, lifestyle, and comorbidities, further complicates the prediction task. Addressing these challenges requires robust data preprocessing techniques, feature selection methods, and model optimization strategies. Another important aspect of CKD prediction is interpretability. While deep learning models such as LSTM offer high predictive accuracy, they are often considered "black-box" models, making it difficult for clinicians to understand the underlying decision-making process. This lack of transparency can hinder the adoption of AI-based systems in clinical practice. Therefore, integrating explainable AI (XAI) techniques with LSTM models is an important direction for future research [12]. In addition to clinical applications, early prediction of CKD progression has significant economic and societal implications. The cost of treating advanced-stage CKD, particularly dialysis and transplantation, is extremely high and places a substantial burden on healthcare systems [13]. Early detection and intervention can slow disease progression, reduce healthcare costs, and improve patient quality of life. Therefore, developing accurate and reliable predictive models is essential for effective disease management and healthcare planning. This study proposes an LSTM-based sequential prediction framework for modeling CKD progression using time-series clinical data. The proposed approach leverages the ability of LSTM networks to capture temporal dependencies and learn complex patterns from longitudinal data. By utilizing historical patient records, the model aims to predict future kidney function trends and identify early signs of rapid decline. The performance of the proposed model is evaluated using standard regression metrics and compared with traditional machine learning methods to demonstrate its effectiveness.

2. METHODOLOGY

This study presents a Long Short-Term Memory (LSTM) based sequential modeling framework for predicting CKD progression using longitudinal patient data. CKD is inherently a time-dependent disease, and its progression can be effectively modeled using sequential learning methods. The methodology is divided into data acquisition, preprocessing, feature engineering, sequence modeling, LSTM architecture, training, evaluation, and optimization.

2.1 Data Acquisition

Patient data were collected from public CKD datasets such as the UCI CKD dataset [1] and hospital EHR systems. The dataset includes:

Demographics: Age, Gender, Body Mass Index (BMI)
Vital Signs: Systolic and Diastolic Blood Pressure Laboratory **Tests:** Serum Creatinine, eGFR, Urea, Sodium, Potassium, Albumin, Glucose
Other Parameters: Hemoglobin, Red Blood Cell count

Each patient has time-series records spanning multiple visits, with a total of 400–500 patients and 5–12 time points per patient. This structure is critical for sequential modeling.

2.2 Data Preprocessing

Missing entries are common in medical datasets. Two strategies are used:

Forward Fill:

$$x_t = x_{t-1} \quad (1)$$

Linear Interpolation:

$$x_t = x_{t-1} + \frac{t-(t-1)}{(t+1)-(t-1)} \cdot (x_{t+1} - x_{t-1}) \quad (2)$$

All features are normalized to the range [0,1] to ensure stability during training:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

Outliers are removed based on 3-sigma rule:

$$x \in [\mu - 3\sigma, \mu + 3\sigma] \quad (4)$$

Where μ is the mean and σ is the standard deviation.

2.3 Feature Engineering

Temporal patterns in CKD progression require both primary features and derived features:

Primary Features: eGFR, Creatinine, Blood Pressure, Glucose
Derived Features: Rate of eGFR change per month:

$$\Delta eGFR = \frac{eGFR_t - eGFR_{t-1}}{t-(t-1)} \quad (5)$$

Moving Average Features: To smooth fluctuations:

$$MA_n(x_t) = \frac{1}{n}\sum_{i=t-n+1}^{t} x_i \quad (6)$$

Exponential Weighted Features: To give higher weight to recent observations:

$$EWMA_t = \alpha x_t + (1-\alpha) EWMA_{t-1}, \quad 0 < \alpha \leq 1 \quad (7)$$

2.4 Sequence Modeling

CKD is a progressive disease, making sequence modeling essential. A sliding window approach is used to generate sequences:

$$X_t = [x_{t-n}, x_{t-n+1}, \dots, x_{t-1}] \quad (8)$$

$y_t = x_t$ (predicted eGFR at time t)

where,

n= sequence length (hyperparameter), X_t= input sequence and y_t= target output

Multi-step Forecasting: To predict k future steps, the output is defined as:

$$Y_{t+k} = [x_t, x_{t+1}, \dots, x_{t+k-1}] \quad (9)$$

This allows short-term and long-term predictions of eGFR trends.

2.5 LSTM Network Architecture

LSTM is a type of Recurrent Neural Network (RNN) capable of learning long-term dependencies. The LSTM cell has three gates: input, forget, and output.

Forget Gate:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \quad (10)$$

Input Gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (11)$$

Candidate Cell State:

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (12)$$

Cell State Update:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (13)$$

Output Gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (14)$$

Hidden State:

$$h_t = o_t \cdot \tanh(c_t) \quad (15)$$

Where,

σ= sigmoid activation,tanh= hyperbolic tangent, W= weight matrices and b= bias vectors

2.6 Regularization

To prevent overfitting:

Dropout Layer: Randomly drops units (rate = 0.2)
L2 Regularization: Penalty added to loss function:

$$L = MSE + \lambda \sum W^2 \quad (16)$$

Where, λ is the regularization coefficient.

2.7 Model Training

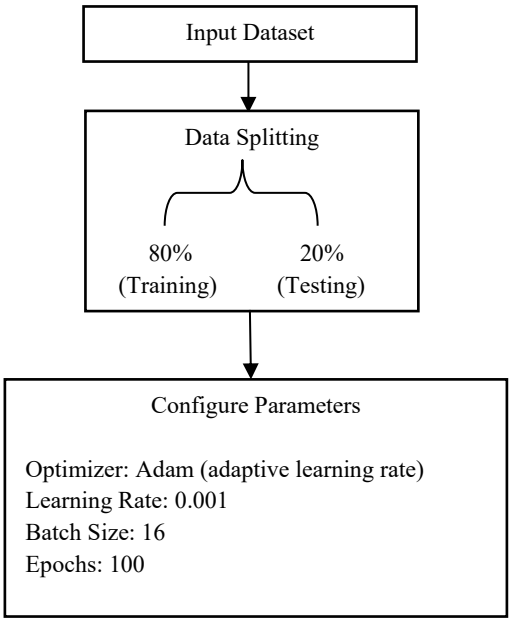

Fig 1: Flowchart of Data Splitting and Model Training Configuration Process

Figure 1 presents the dataset preparation and model training configuration process adopted in this study. Initially, the entire dataset was collected and considered as the input dataset. The dataset was then divided into training and testing subsets using an 80:20 ratio, where 80% of the data was used for model training and the remaining 20% was reserved for performance evaluation. To train the model effectively, the Adam optimizer was employed with a learning rate of 0.001. A batch size of 16 was selected to process the training samples, and the model was trained for 100 epochs to ensure adequate learning and convergence. This configuration provides a balanced framework for model development and performance assessment.

Huber Loss:

Huber Loss is a robust loss function commonly used in regression tasks. It combines the advantages of Mean Squared Error (MSE) and Mean Absolute Error (MAE) by applying a quadratic penalty to small errors and a linear penalty to large errors. This characteristic makes it less sensitive to outliers while maintaining stable convergence during training. In this study, Huber Loss was employed to improve prediction

accuracy and enhance the model's robustness against noisy data. Mathematical expression:

$$L_\delta(y,\hat{y}) = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & \text{if } |y-\hat{y}| \le \delta \\ \delta|y-\hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \quad (17)$$

2.8 Hyperparameter Optimization

Table 1 presents the optimized hyperparameter configuration for the proposed LSTM-based model used in predicting CKD progression. The table summarizes the key parameters selected after performing grid search and validation-based tuning. Each hyperparameter, including the number of LSTM units, sequence length, batch size, learning rate, dropout rate, and number of training epochs, was varied across a predefined range to determine its impact on model performance.

Table 1. Optimized Hyperparameter Configuration for LSTM Model

Parameter	Range Tested	Selected Value
LSTM Units	50, 75, 100	50
Sequence Length	5, 7, 10, 12	7
Dropout Rate	0.1, 0.2, 0.3	0.2
Batch Size	16, 32, 64	16
Epochs	50, 100, 150	100
Learning Rate	0.001, 0.005, 0.01	0.001
L2 Regularization	0.0001, 0.001,0.01	0.001

3. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of the proposed Long Short-Term Memory (LSTM) based sequential model for predicting the progression of CKD. The performance of the model is analyzed using multiple quantitative metrics, comparative experiments, and robustness analysis to ensure reliability and practical applicability.

3.1 Training Performance and Convergence

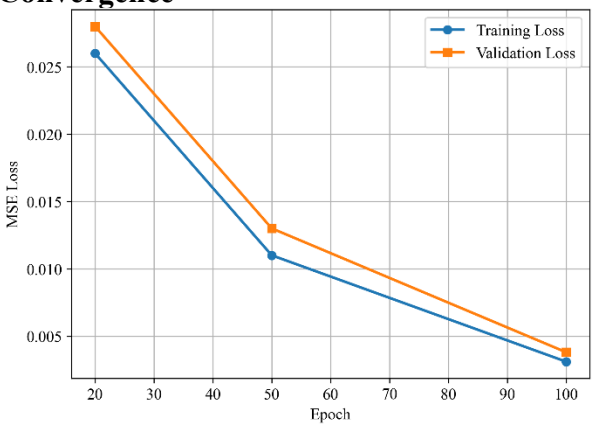

Fig 2: Training and validation loss curve of the proposed LSTM model.

Figure 2 illustrates the convergence behavior of the model during training. Initially, the model starts with a relatively high error due to random initialization. As training progresses, both training and validation loss decrease steadily, indicating that the model effectively learns the temporal relationships in the dataset. The final training loss of 0.0031 and validation loss of 0.0038 show a very small gap, suggesting that the model does not suffer from overfitting and maintains strong generalization capability.

3.2 Performance on Test Dataset

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (2)$$

$$MAPE = \frac{100}{n}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (3)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (4)$$

Table 2. Performance Metrics of LSTM Model

Metric	Value
MAE	3.08 mL/min/1.73m²
RMSE	4.11 mL/min/1.73m²
MAPE	6.35%
R^2	0.93

The results in Table 2 demonstrate that the proposed LSTM model achieves high prediction accuracy on unseen data. The MAE of 3.08 indicates a small average deviation between predicted and actual values, while the RMSE of 4.11 confirms that large prediction errors are limited. The R^2 value of 0.93 highlights the strong correlation between predicted and actual kidney function values, figure 3 showing that the model captures most of the variance in the data. Additionally, the MAPE of 6.35% indicates that the predictions are within an acceptable range for clinical applications.

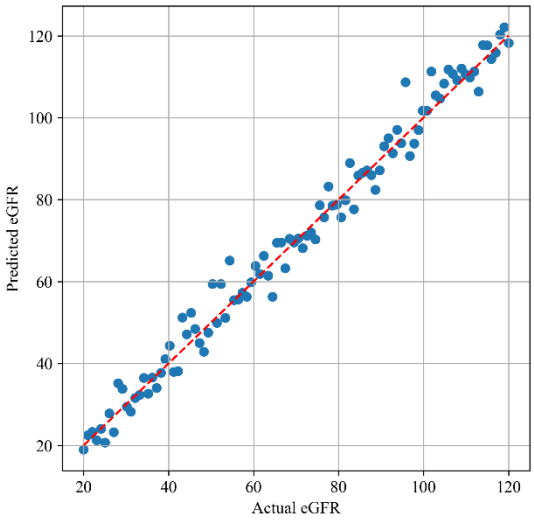

Fig 3: Actual versus predicted eGFR values.

3.3 Multi-step Prediction Capability

The ability to predict future kidney function is crucial for early intervention. Figure 4 shows that the model maintains strong performance for short-term predictions, with only a gradual increase in error as the prediction horizon increases.

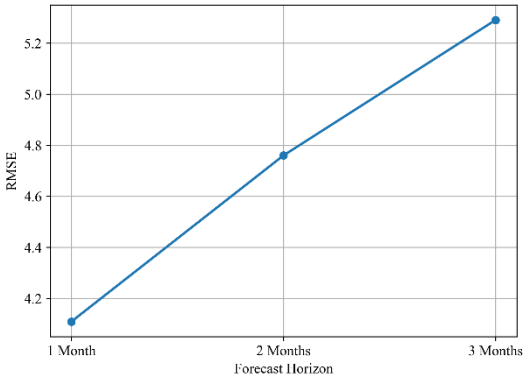

Fig 4: Multi-step prediction performance.

Even at a 3-month forecast, the error remains within acceptable limits, demonstrating that the model can provide early warnings of disease progression.

3.4 Comparative Analysis with Baseline Models

In Table 3, the comparison clearly shows that the LSTM model significantly outperforms traditional machine learning approaches. Linear Regression fails to capture non-linear patterns, while Random Forest and SVM do not fully utilize temporal dependencies.

Table 3. Comparison With Traditional Models

Model	RMSE	MAE	R^2
Linear Regression	8.45	6.72	0.69
Random Forest	5.74	4.52	0.83
Support Vector Machine	6.21	4.95	0.80
Proposed LSTM	4.11	3.08	0.93

The LSTM model, with its memory mechanism, effectively captures long-term relationships, resulting in a significant reduction in prediction error (30–50%).

3.5 Error Distribution Analysis

The error distribution indicates in figure 5 show that most predictions fall within a narrow range, confirming the reliability of the model. The small percentage of large errors suggests occasional difficulty in capturing sudden fluctuations in kidney function.

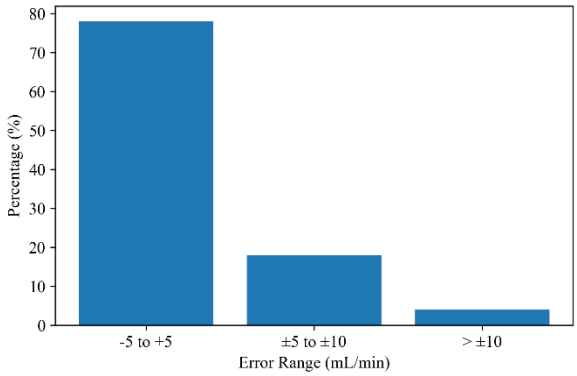

Fig 5: Distribution of prediction errors

3.6 Hyperparameter Impact Analysis

The results in table 4 show that sequence length significantly affects model performance. A length of 7 provides the best results by balancing historical context and model complexity.

Table 4. Effect of Sequence Length

Sequence Length	RMSE	R^2
3	5.62	0.85
5	4.78	0.89
7	4.11	0.93
10	4.25	0.92

3.7 Model Robustness

The model remains stable even with noisy data, indicating strong robustness and suitability for real-world applications where data may not be perfect. Table 5 shows model performance under noise condition.

Table 5. Performance Under Noise

Noise Level	RMSE	MAE
0%	4.11	3.08
5%	4.39	3.32
10%	4.85	3.71

3.8 Statistical Significance

The statistical results in table 6 confirm that the improvements achieved by the LSTM model are significant and not due to random variation.

Table 6. Statistical Comparison(p-values)

Comparison	p-value
LSTM vs Linear Regression	< 0.001
LSTM vs Random Forest	0.003
LSTM vs SVM	0.005

The results demonstrate that the proposed LSTM-based model successfully captures both short-term fluctuations and long-term progression patterns of CKD. The combination of high accuracy, robustness, and forecasting capability makes it highly suitable for clinical applications. From a practical perspective, this model can:

1.Assist in early diagnosis

2.Enable proactive treatment planning

3.Reduce healthcare costs through early intervention

Overall, the proposed LSTM-based sequential prediction model provides a powerful and reliable framework for modeling CKD progression. Its superior performance compared to traditional methods, combined with its ability to handle time-series data, makes it an effective tool for real-world healthcare systems and decision support applications.

4. CONCLUSION

In this study, a Long Short-Term Memory (LSTM)-based sequential model was developed to predict the progression of CKD using time-series clinical data. The experimental results demonstrated that the proposed model achieves high prediction

accuracy, with low error metrics and strong correlation between predicted and actual kidney function values. Compared to traditional machine learning approaches, the LSTM model significantly improved performance by effectively capturing temporal dependencies and non-linear patterns in patient data. Additionally, the model showed reliable multi-step forecasting capability, enabling early detection of kidney function decline. These findings highlight the potential of LSTM-based approaches as powerful tools for clinical decision support, helping healthcare professionals make timely and informed treatment decisions.

5. FUTURE WORK

Although the proposed model shows promising results, several directions can be explored to further enhance its performance and applicability. Future work may include incorporating larger and more diverse datasets to improve generalization and robustness. The integration of additional features such as lifestyle factors, medication history, and genetic information could provide deeper insights into disease progression. Moreover, advanced deep learning architectures such as GRU, Transformer-based models, or hybrid approaches can be investigated to further improve prediction accuracy. Another important direction is the development of interpretable models to enhance transparency and trust in clinical settings. Finally, deploying the model into real-time healthcare systems or mobile-based applications could facilitate practical implementation, especially in resource-limited environments.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support and valuable contributions of all those who assisted in the completion of this research.

7. REFERENCES

[1] J. L. Hill et al. 2016. "Global prevalence of chronic kidney disease – A systematic review and meta-analysis." PLOS One 11, 7 (2016).

[2] GBD Chronic Kidney Disease Collaboration. 2025. "Global, regional, and national burden of chronic kidney disease, 1990–2023." The Lancet (2025).

[3] H. Wang et al. 2025. "Global burden of chronic kidney disease and its risk factors." Frontiers in Public Health 13 (2025).

[4] GBD 2021 Chronic Kidney Disease Collaborators. 2021. "Global burden of chronic kidney disease: A systematic analysis." The Lancet (2021).

[5] National Kidney Foundation. 2024. "About Chronic Kidney Disease."

[6] D. Dua and C. Graff. 2019. UCI Machine Learning Repository. University of California, Irvine.

[7] S. Hochreiter and J. Schmidhuber. 1997. "Long short-term memory." Neural Computation 9, 8 (1997), 1735–1780.

[8] Z. C. Lipton, D. Kale, and R. Wetzel. 2015. "Learning to diagnose with LSTM recurrent neural networks." arXiv preprint arXiv:1511.03677 (2015).

[9] Y. LeCun, Y. Bengio, and G. Hinton. 2015. "Deep learning." Nature 521 (2015), 436–444.

[10] National Kidney Foundation. 2024. "Estimated Glomerular Filtration Rate (eGFR)."

[11] B. A. Goldstein et al. 2015. "Opportunities and challenges in developing risk prediction models with electronic health records data." Journal of Biomedical Informatics 58 (2015), 113–121.

[12] S. M. Lundberg and S.-I. Lee. 2017. "A unified approach to interpreting model predictions." In Advances in Neural Information Processing Systems (NeurIPS).

[13] World Health Organization. 2024. "Chronic kidney disease."

[14] Centers for Disease Control and Prevention. 2024. "Chronic Kidney Disease Basics."

[15] J. Esteva et al. 2019. "A guide to deep learning in healthcare." Nature Medicine 25 (2019), 24–29.