Advances in Computational Design, *Vol. 11, No. 1 (2026) 63-77*
https://doi.org/10.12989/acd.2026.11.1.063

Techno-Press

Electronic health records (EHRs) in chronic kidney disease classification using LSTM

Pradeep Balaji B.*[1], Gayathri M.[2], Deepika Sirmoria[3a], Job Prasanth Kumar Chinta Kunta[4]

[1]*Technical Specialist Network – Security, Tech Hat Pvt Ltd, Bangalore, Inida*
[2]*Operations Head, Willron Electronics, Bangalore, India*
[3]*Computer Science Engineering, Anurag University, Ghatkesar, India*
[4]*Lead Solution Data Architect, The Automobile Association, London, United Kingdom*

(Received September 30, 2025, Revised October 19, 2025, Accepted October 20, 2025)

Abstract. Chronic Kidney Disease (CKD) is among the most significant global health concerns, particularly in terms of its insidiousness during the first stage of its development and gradual devastation throughout the years. There is a prospect of utilizing Electronic Health Records (EHRs) to improve the outcome due to the ability to address problems at an early stage to deliver the most efficient intervention. The paper presents an intelligent predictive analytics system of healthcare in Abu Dhabi healthcare systems that is built on the EHR data collected. The pipeline of the framework is systematic and it entails data preprocessing, feature extraction and classification. The preprocessing phase is assigned to aligning the data, its coherence, and the removal of redundancies and the handling of missing values across all the EHR datasets. The step is relevant due to the heterogeneous nature of clinical information being rather complex. In summarizing the data, Principal Component Analysis (PCA) is applied to extract the features by subjecting the data to the process to compress the data and retain the most clinical information. This improves the computational and model efficiency and performance by removing noise and redundancy. It is then inputted into the constructed Long Short-Term Memory (LSTM) network due to its learning capabilities which give long-range dependencies and temporal patterns of sequential patient information. Precision, recall and F1-score, are also used to test the effectiveness of the model by determining whether the model is effective in the proper identification of CKD cases. The findings show that LSTM model is better than the traditional classifiers it is more predictive and robust. As highlighted in this paper, advanced deep learning methods might be used on EHR data to aid in the prompt identification of CKD and enhance the clinical decision-making process. The suggested framework is flexible and can be extended and provide useful information on how the framework can be applied in real-life healthcare.

Keywords: accuracy; CKD; F1-score; harmonization; HER; LSTM; precision; recall

1. Introduction

High blood pressure, renal failure, congestive heart failure (CHF), and Type 2 Diabetes Mellitus (110) are among the causes of chronic illnesses. In the modern world, these are some of

*Corresponding author, Technical Specialist, E-mail: dikuma0070@gmail.com
[a] Ph.D., E-mail: deepikasirmoria9963@gmail.com

ISSN: 2092-7614 (Print), 2092-7622 (Online)

the causes of illness and death. According to the World Health Organisation (WHO) 71 percent of all deaths in the world are because of chronic diseases and most of these diseases are preventable or treatable when they are detected early in life [1]. Unluckily, it is through the insidious and gradual development of symptoms, lack of access to healthcare services and late clinical tests that most patients are not diagnosed at an early stage [2]. These diseases develop in people over years with little obvious signs till clinical symptoms appear, and therefore the early diagnosis becomes desirable. One of the most appropriate methods that assist patients in achieving their better outcomes, paying less on medical care, and leading more active lives is the early diagnosis of chronic diseases. By taking broader adoption of electronic health records (EHRs), health care professionals can take a treasure trove of patient information which can subsequently be mined to deduce clues that border these diseases [3]. The research on which this project was based however identified that the bulk and abundance of EHR data create crude analytical problems that cannot be dealt with through standard methods. Machine learning is one of the fields of AI that has lately been developed into its deep learning to enhance the analysis of EHR. Unlike other types of machine learning algorithms, which rely on hand crafted features, deep learning models are learned using the raw data [4]. This feature is the reason why deep learning can be useful when dealing with large and non-structured datasets like EHRs and the information they contain that on top of being numerically coded (e.g., lab results and medication prescription) may also comprise free text, clinical notes, and imaging reports; i.e., structured, numerically coded, and unstructured data [5].

Use of Electronic Health Record (EHR) system in the medical sector has been on the rise in most countries around the world in the past few years. The EHRs are more efficient, manageable, and become more accessible compared to the traditional paper records. This increased usage has seen a significant rise in the amount of medical data being produced. Even though EHRs were initially designed to store the data, the researchers have found that the abundance of the information stored in them could prove extremely useful in different clinical informatics applications [6]. EHR data were firstly analysed using conventional machine learning methods. The approaches have however not been enough to enable sound and scalable analysis. However, an alternative solution to the problem of handling and analysing large-scale medical data is presented by the rapid development of deep learning technologies that are backed by strong computing power. Deep learning models, especially those that consider features based on complex data, have already proven to be able to present accurate and meaningful information in the real-world clinical setting. Due to the development of research in the given sphere, the use of deep learning on EHR data has grown in four major directions, including representation learning, medical prediction, privacy protection, and information extraction. The growing adoption of deep learning in EHR-related applications is not only supplanting conventional approaches to analysis but also showing performance of an expert in performing in a variety of medical tasks [7].

1.1 Contributions

- This paper suggests a new end-to-end predictive model to identify chronic kidney disease at an early stage with the help of real-world Electronic Health Records (EHRs) of the Abu Dhabi healthcare system.
- The paper proposes a powerful data harmonization pipeline to clean, standardize, and unify heterogeneous EHR data, which guarantees uniformity and consistency over patient records. PCA- It uses PCA to identify and store the most informative clinical features, which result in

computational efficiency and less noise in high-dimensional EHR data.

- LSTM model is used to learn intricate temporal associations and sequential tendencies in patient health records that more often than not are not covered by traditional models.
- The framework is strictly justified by means of precision, recall and F1-score, which prove high performance of the LSTM model in case of CKD cases over the baseline classifiers.
- The suggested solution is flexible to most EHR systems and can be incorporated into clinical processes to assist in decision-making, early diagnosis, and resource optimization in medical organizations.

1.2 Objectives

- To collect and utilize real-world EHR data from Abu Dhabi healthcare systems for the identification and classification of CKD cases.
- To perform data preprocessing through harmonization techniques that standardize and clean heterogeneous clinical data, ensuring consistency and quality across patient records.
- To improve computational efficiency and preserve the most pertinent medical features by using PCA for feature extraction and dimensionality reduction.
- To put into practice a LSTM model that can accurately classify CKD patients by identifying temporal and sequential patterns in EHR data.

Section 2: Literature Review presents a comprehensive overview of recent advancements in chronic kidney disease (CKD) prediction, EHR-based analytics, and machine learning and deep learning methodologies. Section 3: Proposed Methodology outlines the overall approach and design of the proposed model. Section 4: Results and Discussion details the model's performance, including metric comparisons, ROC curves, and confusion matrices. Finally, Section 5: Conclusion summarizes the key findings of the study and highlights its major contributions.

2. Related works

[8] The study will develop a categorisation of patient safety incident reporting that occurred immediately after the implementation of a new EHR system based on a data set of the occurrence of these events. The categorisation scale was created with the help of Finnish Technology-Induced Error Risk Assessment Scale instrument. The study discovered that additional mistake categories were added to the developing taxonomy, and that the frequency of patient safety events using EHRs grew fivefold throughout the deployment period. The most frequent mistakes were related to clinical process, documentation, usability, and interface issues. According to the study, reports with inadequate information were rejected in 14.8% of cases.

[9] The goal of the project is to create a machine learning-based, interoperable electronic health record (EHR) system that will help diagnose diabetes early. Using a decision support system, the database structure and models for data extraction, cleaning, and processing were evaluated on the medical records of 1080 patients. To enable early diabetes prediction, patient health records will be gathered, stored, and shared.

[10] Even in public health institutions, medical records especially in poor countries such as India, it is still hard to digitise. There is an increasing use of healthcare IT tools in some institutions like Electronic Medical Records (EMR) and Hospital Information Systems which are

proprietary and could not be shared across hospitals. In order to streamline the processes within health services and enable the efficient and continuous treatment of patients, a proposed secure Electronic Health Record (EHR) architecture is developed based on the use of accepted medical language and coding standards.

[11] The paper suggests that the Internet of Things (IoT)-based devices and EHRs can be used as a predictor of cardiac disease. It is also based on cluster primal-dual splitting architecture, federated learning architecture, and soft-margin L1-regularized SVM classifier. This type of team-based predictive analytics system improves patient outcomes and data security more effectively and increases participation and compliance with regulations.

[12] It is suggested that the Hierarchical Autoregressive Language model (HALO) is the most effective method for creating high-fidelity synthetic EHRs as it guarantees the statistical characteristics of actual EHRs and enables precise training of machine learning models without running the risk of privacy concerns. Experiments show that HALO can provide high-fidelity data, which increases prediction accuracy since illness codes are likely to be extremely similar to the actual EHR data.

[13] The article provides a distinctive solution to the natural language processing (NLP) in the medical context with the help of deterministic rules to extract control information at the fine-granularity, semi-supervised learning of limited clinical vocabulary, and unsupervised learning of word representations. This language-neutral approach improves the clinical decision support performance, interpretability, and transparency.

[14] EHR systems propagate mistakes, encourage copy-paste, and increase effort. Errors occur in 15% of cancer diagnostic and therapy charts. Accuracy may be increased with the use of engagement tools, patient portals, and modern technologies.

[15] The study developed a model for identifying and classifying illness trends in electronic health records (EHRs) using data from emergency department visits and samples from the National Hospital and Ambulatory Medical Care Survey. AUROC and t-SNE were used to evaluate the model's performance, and BERT, Deep InfoMax, and SimCLR were used to include sickness concepts.

[16] Healthcare data management has been profoundly changed by digital technology, yet issues with processing, storage, security, privacy, and usability still exist. Although blockchain technology and natural language processing are the main subjects of studies, problems with governance regulations, distrust, scalability, security, privacy, poor performance, and excessive prices still exist.

[17] In this work, regular clinical data from the Wisconsin Sleep Cohort dataset is used to identify OSA using machine learning techniques. The model separates OSA patients from non-OSA patients based on characteristics such patient demographics, blood reports, physical measures, and sleep history. The model obtained exceptional evaluation ratings with a sensitivity of 88.76%, specificity of 40.74%, F1-score of 75.96%, accuracy of 68.06%, PPV of 66.36%, and NPV of 73.33%.

[18] In contrast to COG's stated laboratory adverse event rates, ExtractEHR accurately detected laboratory adverse event rates for juvenile acute leukaemia. These rates can be used to compare various treatments and to inform patients about the risks associated with chemotherapy. ExtractEHR AE ascertainment is a novel approach to improving laboratory AE trial reporting.

[19] Using electronic health information, the Spanish Sistema Nacional de Salud created a bioinformatics prediction system to forecast inadequate treatment of chronic respiratory conditions including COPD and asthma. Although the model's sensitivity, specificity, and accuracy were

excellent, further proof is needed.

[20] The study demonstrates that the University of California, San Francisco's electronic medical data and knowledge networks may be used to contextualise sex dimorphism, prioritise biological theories, and anticipate the start of Alzheimer's disease. The random forest models and matched cohort models were used in the study to determine the conditions that had predictive power prior to the onset of AD. Genetic analysis of colocalization helps support the association of AD with hyperlipidemia and more intense female association of AD with osteoporosis.

[1] The article proposes a deep learning model in terms of Long Short-Term Memory networks and attention to detect the patterns of early onset of chronic diseases such as T2DM, hypertension, CKD, and CHF. The model was superior to the traditional models and was found to be demographically sound. Its greatest performance was achieved in T2DM by attaining 92.0% F1-Score. The ability to generalize to domain changes was seen in generalization experiments on MIMIC-III and eICU datasets by using the framework. Further extensions will be on integration.

[21] An artificial intelligence prognosis model for heart failure in diabetics has been developed. Patients having a clinical examination of cardiology and no history of heart failure diagnosis were included in the sample of this retrospective cohort research, which was based on electronic health records (EHR). The Cox proportional hazard model and deep neural network survival were used to develop two prognostic models. Twenty predictors of different domains that reflect changes in clinical practice were discovered by the AI technique. The results suggest that EHRs and AI techniques for survival analysis might improve diabetic heart failure prognostic models.

[22] This study addresses performance metrics and data preparation concerns while investigating deep learning models for diabetes prediction using EHR data. While highlighting recent advancements such as explainable AI, wearable health device integration, and federated learning, it also urges more study to increase model robustness.

[23] In order to forecast clinical deterioration in high-risk patients, the study created a high-dimensional model utilising EHR data. By forecasting 1028 EHR characteristics using the Intensive Care Warning Index (I-WIN), doctors may focus on treatments. The study suggests using machine learning-based risk prediction in place of expert consensus.

[24] In order to forecast the development of chronic kidney disease (CKD) from stage II/III to stage IV/V, the authors created a pipeline to assess electronic health information and construct recurrent neural network models, which had an AUROC of 0.957.

[25] Deep learning is being used to analyse Electronic Health Records (EHR) and integrate it with blockchain which is a decentralised, immutable, shared, and distributed ledger. The work will allow a deep learning algorithm to process EHR data on the blockchain and warn patients with consultations and diet plans. The integrated environment will enhance security of data and minimize intervention of third parties.

[26] In retrospective case-control research conducted at the University of Missouri Healthcare, machine learning models were created and verified to identify and forecast early Alzheimer's disease in de-identified electronic health record data. 380, 269 people 40 years of age and older who had two or more medical visits were used by the researchers. The GBT model achieved the greatest AUC-ROC scores out of the six ML classifier models that were evaluated. The study highlights how ML models may be used with EHR data to improve patient outcomes, provide prompt intervention, and predict ADRD early.

[27] Researchers use GFR and indicators of kidney damage to identify CKD that can increase the chances of premature mortality. An original deep learning model has been developed to predict and early identify CKD. The model selects the key features via Recursive Feature Elimination and

then machine learning models are provided with the features. The model can be employed by nephrologists as it has 100 percent accuracy.

[28] The study explores the use of machine learning methods in the prediction of chronic kidney disease using a dataset and seven classifier algorithms. The study found that penalty L2 of LSVM performed best and its accuracy is 98.86.

[29] To be able to manage this and reduce the number of health problems, it is necessary to diagnose CKD at its initial stages. This study employs machine learning techniques of stage prediction by the use of the Random Forest, Support Vector Machine, and Decision Tree models. The results show that recursive elimination feature-based RF is superior to SVM and DT.

[30] Unlike traditional methods, this article uses optimised neural networks to diagnose chronic kidney disease (CKD) with the help of the UCI machine learning dataset. The paper compares optimised CNN, ANN and LSTM models by binary classification based on 24 characteristics.

3. Proposed methodology

The proposed resolution is a predictive healthcare analytics system, which is aimed at permitting the early detection of CKD on the ground of Electronic Health Records (EHRs) received with the Abu Dhabi healthcare system. The four main building pieces of the dataset are the pre-processing, feature extraction, and classification. To ensure that patient records are interoperable with other healthcare formats, the raw EHR data should be first of all collected and later be subject to data harmonisation procedures. After the pre-processing stage, the data undergoes the Principal Component Analysis (PCA), which assists in improving the computational and model efficiency by extracting the most significant clinical characteristic and reducing them into a smaller set of clinically significant variables. The complex feature collection is then categorised using a LSTM network. Since it has the ability to identify these complex temporal correlations and sequential patterns in longitudinal healthcare data, which most machine learning-based models fail to do, the LSTM model was selected. The system is also trained and tested using the harmonised EHR dataset, and key performance indicators including accuracy, precision, recall, and F1-score are used to gauge the model's effectiveness.

3.1 Dataset: Chronic kidney disease EHRs Abu Dhabi

Between January 1st and December 31st, 2008, 491 patient computerised medical records were sampled at the Tawam Hospital in Al-Ain, Abu Dhabi, United Arab Emirates. There were 241 female patients and 250 male patients, with an average age of 53.2 years. All the required information (laboratory tests results, examination results, and medical history) will be entered by each patient on the chart that has 22 clinical factors. The patients that were considered in this research according to the criteria of Tawam Hospital had either cardiovascular disease or were at risk of developing it.

3.2 Preprocessing: data harmonization

The process of integrating and standardising electronic health data from several systems and sources into a uniform, standardised format is known as EHR data harmonisation. This makes it possible for dependable healthcare delivery across systems, accurate analysis, and interoperability.

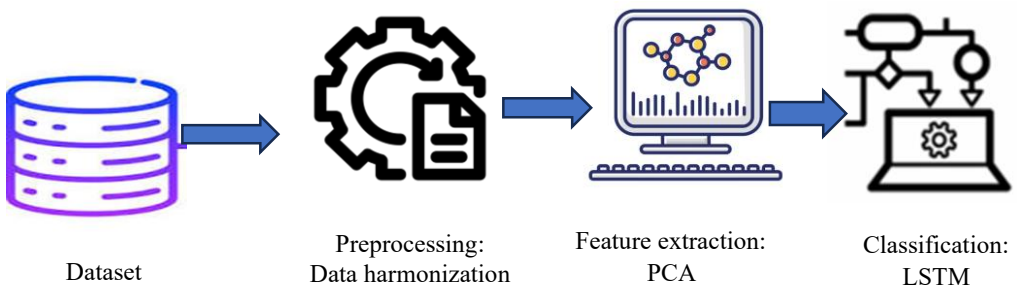

Figure 1. Flow diagram of HER for CKD

The provision of EHR datasets is

$$D_i = \{x_{i1}, x_{i2}, \dots \dots \dots, x_{in}\} \tag{1}$$

H: Harmonization function that transforms raw EHR data into a standard format

$D*$: Final harmonized EHR dataset

$$D^* = \bigcup_{i=1}^{k} H(D_i) \tag{2}$$

This means we apply harmonization to each dataset D_i, and then combine them into a unified dataset D∗

3.3 Feature extraction: PCA (principal component Analysis)

The PCA feature extraction method of data reduction downsizes the large data set in high dimensions to a small one by identifying the directions within the data that have the largest variance. The first step will be to standardise the dataset to ensure that all the features bring equal contribution to the analytical process and get calculated by the algorithm. Mean and standard deviation of every attribute are computed to alter the information. This is followed by covariance to establish the connections between the features collection. The eigenvectors and eigenvalues are obtained. The overall variance which each eigenvector accounts is represented in the resultant eigenvalues. Accordingly, in this eigenvalue computation, the collection of eigenvalues that is more likely to describe most of the patterns that are observed is selected according to their sizes, relative to their largest. Based on the projection of their data upon these principle components, the most significant information of the data is obtained in a reduced dimensional form. It is among the PCA remedies to overcome those high-dimensional data issues, especially to make the machine learning models more knowledgeable and affordable in the context of processing expenses. The features are provided by extracting the FSs with the help of PCA. The numbers of the input will now be standardised with the statistical calculations like the Mean (M), Standard Deviation (SD) and the Covariance (Cv).

$$Mean = \frac{x_m}{m}, \tag{3}$$

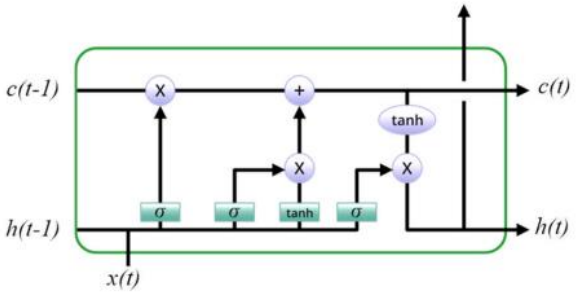

Figure 2. LSTM architecture

Here $x_m - data\ match, x - total\ number\ of\ data$

$$SD = \sqrt{\frac{\sum(y_i - M)^2}{N}}, \tag{4}$$

$$y_i - value\ from\ dataset, M\ is\ the\ mean\ and\ N - size$$

$$CV = \begin{pmatrix} \delta_{x_1x_1} & \cdots & \delta_{x_1x_n} \\ \vdots & \ddots & \vdots \\ \delta_{x_nx_1} & \cdots & \delta_{x_nx_n} \end{pmatrix}. \tag{5}$$

Standard CV is used to indicate the covariance between x and y characteristics, followed by the eigen vector ev1, ev2, … evn and eigen values $\lambda_1, \lambda_2, \ldots\ldots \lambda_n$

$$ev1^T cvev_i = ev_i^T(\lambda_i(\| evi \|^2))^2 = \lambda_i \tag{6}$$

LSTM (Long short-term memory)

To learn and predict series, other researchers have used deep LSTM RNN architecture. Unlike regular RNNs, which may also experience the issue of vanishing gradients in training with a long sequence, LSTMs are programmed to identify long-term relationships. Due to the engineering of the LSTMs with long-term storage capacity memory cells, the network is capable of recalling and utilizing important information in the past time-steps. Furthermore, LSTMs have input, forget, and output gates that control the flow of information. Since the aforementioned gates control how information enters and leaves the memory cells, the LSTM may keep or delete information based on its relevance. Fig. 2 depicts LSTM in its general form.

The forget gate, also known as fg(t), is used at the beginning of the LSTM network operation to decide which data from the previous state the memory cell unit should discard. Eq. (7) represents the forget gate, fg(t).

$$f_g(t) = \sigma(\alpha_{f_g} x(t) + \beta_{f_g} h(t-1) + \delta_{f_g}) \tag{7}$$

The forget gate permits values in the range of 0-1, as shown in the equation $f_g(t)$. It uses the logistic sigmoid function, σ, α_{f_g} β_{f_g}, and δ_{f_g} the term that can be used to represent the biasing vector and adjustable weighting matrices. Then the next step involves identifying what should be typed in the memory cell unit to update it. The sigmoid function selects the input gate, i(t), which

determines the values that should be updated. There is also a hyperbolic tangent (tanh) layer that is used in the event of a possible update vector that is set to be C (t). The details of calculating i(t) and C(t) are explained by use of Eqs. It is absent from Eqs. (8) and (9):

$$i(t) = \sigma(\beta_i h(t-1) + \alpha_i x(t) + \delta_i) \tag{8}$$

$$c(t) = \tanh\left(\beta_c h(t-1) + \alpha_c x(t) + \delta_c\right) \tag{9}$$

It is a vector that takes on values between 0 and 1 in the equations. While αc, βc, and δc refer to a collection of trainable parameters, αi, βi, and δi represent a set of trainable parameters associated with the input gate.

$$c(t) = i(t)^{\circ}c(t) + f_g(t)^{\circ}c(t-1) \tag{10}$$

The ∘ symbol represents element-wise multiplicity. 0 c(t) indicates fresh data that will be added to the cell state, whereas g(t) 0 c(t 1) represents data that has been collected over time and will ultimately be discarded. Calculating the output gate is the final step that is crucial to determining the hidden state, h (t). This will be accomplished by computing the output, denoted as o(t), using the sigmoid function. Nevertheless, the outcome is obtained by multiplying the hyperbolic tangent output by the output of o(t). The formulae used to illustrate the computation are shown in Eq. (11) and (12).

$$O(t) = \sigma(\alpha_o x(t) + \beta_o h(t-1) + \delta_o) \tag{11}$$

$$h(t) = O(t)^{\circ}\tanh\left(c(t)\right) \tag{12}$$

In this equation, O(t) represents a vector [0,1], where αo, βo, and δo stand for trainable input gate parameters

4. Result and discussion

The proposed Electronic Health Records (EHR) system is based on predictive healthcare analytics of chronic renal disease. LSTM is the categorisation model that is effective in managing and interpreting medical data. The system will be developed and the data analysed safely and effectively with the help of Python 3.11, that is compatible with a number of AI and machine learning libraries. The LSTM model processes EHR data and uses significant performance indicators, including the training loss and prediction accuracy. Besides making medical predictions more accurate, this would protect the patient information hence leading to improved and more certain healthcare results.

Fig. 3 illustrates that the LSTM model demonstrates strong accuracy in classifying EHR data, successfully generating reliable predictions for chronic diseases. This performance reflects the model's capability to recognize meaningful patterns within large, high-dimensional clinical datasets. By leveraging patient demographics, laboratory results, and medical history features commonly available in EHRs the LSTM classifier supports early diagnosis and clinical decision-making. Furthermore, the model exhibits strong stability against overfitting, making it suitable for real-world healthcare applications and adaptable across diverse EHR systems.

The training and validation losses of the LSTM based model as shown by Fig. 3 indicate the learning curve and training behaviour of the model. The trend of the reduction of the values of

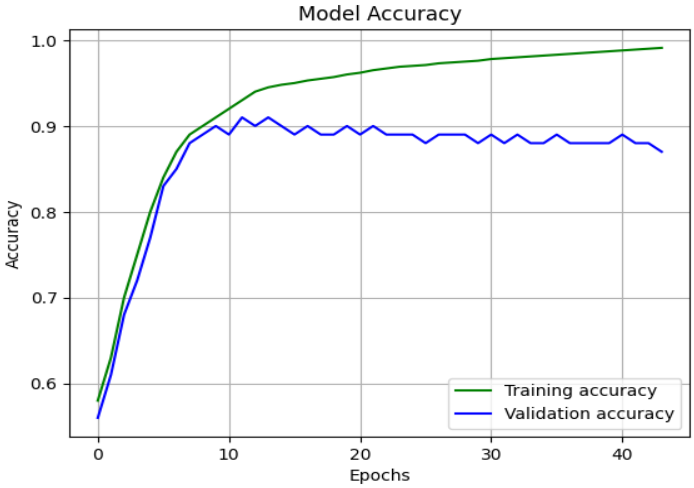

Figure 3. Accuracy of HER for chronic disease

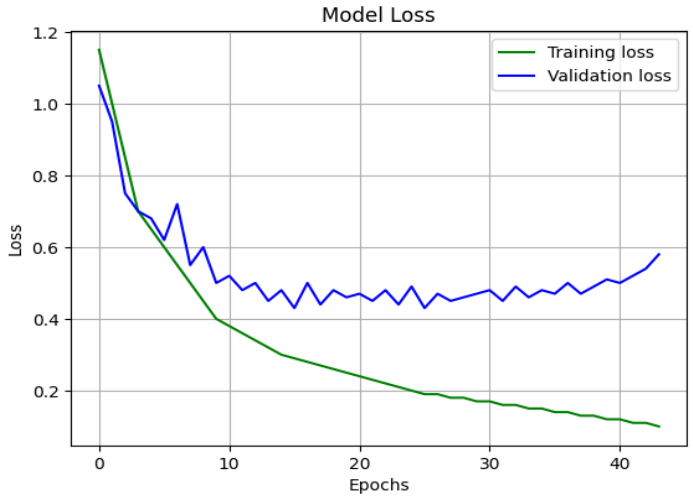

Figure 4. Losses of HER for chronic disease

both the losses over time show that the optimization process is successful and the model is learning meaningful patterns on the EHR data concerning chronic kidney disease. In addition, the fact that there was no serious overfitting also supports the fact that the model can be generalized well to unobserved data, which improves its availability in the real-world clinical setting.

4.1 Performance metrics

Correct classification of CKD samples is indicated by a high TP value. Conversely, FN represents CKD cases that were misclassified. False Positives (FP) indicate notCKD samples that were incorrectly predicted as CKD, while True Negatives (TN) confirm correct classification of notCKD samples.

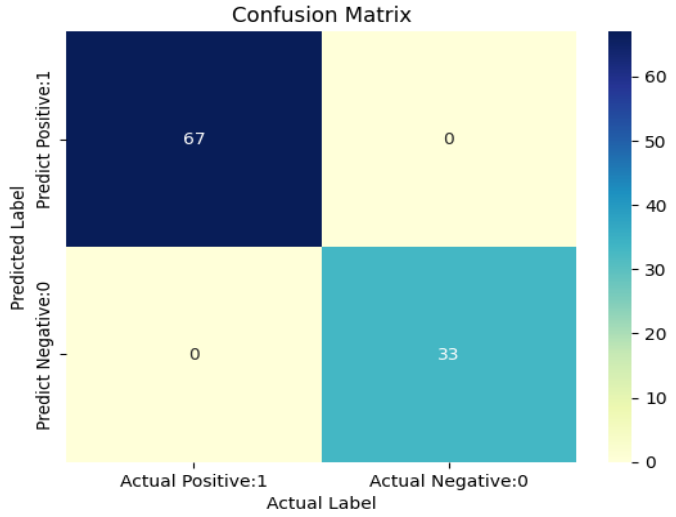

Figure 5. confusion matrix

Accuracy

It is the proportion of accurate estimates to all forecasts. The capacity to accurately forecast a circumstance is one definition of accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Recall

Recall is the proportion of correctly revealed positive observations to all observations in the category, as the following equation illustrates.

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

Precision

This measure represents the proportion of accurate positive predictions among all predictive positive observations, as shown by the equation below.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

F1-score

In the F-measure, precision and recall are weighted averages. There are both false positives and false negatives in the process. What the term "F-measure" means is

$$F1 - score = \frac{2 * precision \times Recall}{Precision + Reall} \tag{16}$$

Here, we provide the suggested model's findings. Chronic kidney disease (CKD) data are made up of 25% testing data and 75% training data. As can be seen in the confusion matrix below Figure 5, the model operated perfectly on the test set, correctly identifying each instance of a true positive and true negative.

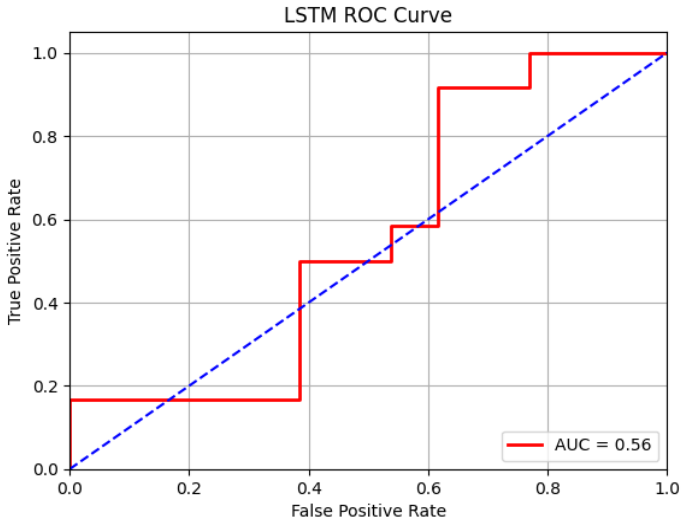

Figure 6. ROC-AUC curve for CKD

Table 1. Comparative analysis

Models	Accuracy (%)	Precision (%)	Recall (%)	F1- Score (%)
KNN	92	98	88	92
SVM	92	96	87	92
LST M	98	97	95	94

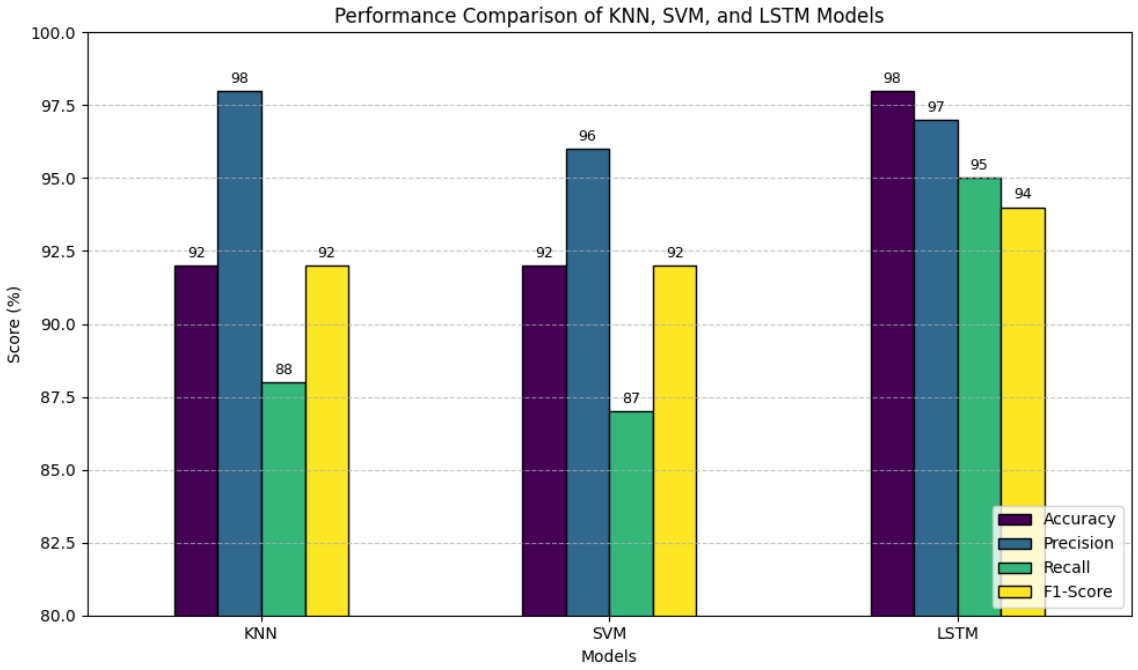

Figure 7. Performance comparison of HER data for CKD

The ROC curve defines the Area Under the Curve (AUC), which is a key metric for evaluating a model's classification performance. An AUC score closer to 1.0 indicates excellent predictive accuracy and a strong ability to distinguish between classes. As demonstrated in Figure 6, the proposed model achieved a near-perfect AUC score, confirming its high effectiveness in correctly identifying chronic kidney disease cases from the EHR data.

The main evaluation criteria for comparing the performance of EHR data for chronic kidney disease (CKD) were the F1-score, accuracy, precision, and recall. The results demonstrate that on all metrics, the LSTM model outperforms the SVM and KNN classifiers. The LSTM model's remarkable ability to recognise complex temporal patterns in EHR data is seen in Table 1 and Fig. 7, which boosts the model's effectiveness for early CKD recognition and clinical decision support.

5. Conclusions

The research proposes a good and clever forecasting algorithm in the early diagnosis of CKD based on the EHRs of the Abu Dhabi healthcare environment. The system guarantees clean and consistent pertinent input to classification through a mix of the most current data pretreatment methods that encompass data harmonisation and dimensionality reduction with the help of Principal Component Analysis (PCA). The conventional machine learning algorithms may not be able to capture the temporal and sequential trends in patient health data, but an LSTM model can be trained to capture these trends effectively. The main performance measures applied to the proposed model were accuracy, precision, recall, and F1-score as 98%, 97%, 95% and 94%. These findings indicate that LSTM model is more effective than these traditional methods because it has high prediction accuracy and generalisability. The fact that the model is able to identify CKD patients accurately is further proven by the visual aids like confusion matrices and ROC curves. To sum up, the developed LSTM-based solution offers a high scalability and efficiency of early CKD diagnosis, and it can be of great use in clinical decision-making. Not only is the framework increasing the predictive accuracy but it also provides the potential of deep learning in helping to transform EHR data into actionable information. Certainly, more work can be done in the future to attempt integration with real-time clinical systems, multi-disease prediction models, and the use of other forms of data (including medical imaging or genetic data).

References

1. Jamal, A., Kumar, P.A., Ampavathi, A., Barot, K.K., Golla, K., Bhosale, Y.H. (2025). deep learning for early diagnosis of chronic conditions using electronic health records. Journal of Neonatal Surgery, 14(18s), 1099-1110.
2. Smith, J., Doe, A. (2020), Deep learning for chronic disease prediction using EHRs. Journal of Medical Systems, 44(3), 1-12.
3. Brown, L., Kim, S. (2021), CNNs for early detection of diabetes from EHR data. Health Informatics Journal, 27(2), 124-135.
4. Zhang, Y., Lee, M. (2019), Recurrent neural networks in chronic disease management. IEEE Transactions on Neural Networks, 30(7), 2102-2113.
5. Patel, R., Singh, T. (2020), Transformers in healthcare: Chronic disease detection. Computers in Biology and Medicine, 123(5), 104021.
6. Xu, J., Xi, X., Chen, J., Sheng, V.S., Ma, J., Cui, Z. (2022). A survey of deep learning for electronic

health records. Applied Sciences, 12(22), 11709. https://doi.org/10.3390/app122211709
7. Poongodi, T., Sumathi, D., Suresh, P., Balusamy, B. (2020). Deep Learning Techniques for Electronic Health Record (EHR) Analysis. In Bio-inspired Neurocomputing, 73-103. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-5495-7_5
8. Palojoki, S., Saranto, K., Reponen, E., Skants, N., Vakkuri, A., Vuokko, R. (2021). Classification of electronic health record–related patient safety incidents: development and validation study. JMIR medical informatics, 9(8), e30470. https://doi.org/10.2196/30470
9. HL, G., Ravi, V., Almeshari, M., Alzamil, Y. (2023). Electronic health record (EHR) System development for study on ehr data-based early prediction of diabetes using machine learning algorithms. The Open Bioinformatics Journal, 16(1). https://doi.org/10.2174/18750362-v16-e230906-2023-15
10. Pai, M.M., Ganiga, R., Pai, R.M., Sinha, R.K. (2021). Standard electronic health record (EHR) framework for Indian healthcare system. Health Services and Outcomes Research Methodology, 21(3), 339-362. https://doi.org/10.1007/s10742-020-00238-0
11. Bebortta, S., Tripathy, S.S., Basheer, S., Chowdhary, C.L. (2023). Fedehr: A federated learning approach towards the prediction of heart diseases in iot-based electronic health records. Diagnostics, 13(20), 3166. https://doi.org/10.3390/diagnostics13203166
12. Theodorou, B., Xiao, C., Sun, J. (2023). Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. Nature communications, 14(1), 5305. https://doi.org/10.1038/s41467-023-41093-0
13. Berge, G. T., Granmo, O.C., Tveit, T. O., Ruthjersen, A.L., Sharma, J. (2023). Combining unsupervised, supervised and rule-based learning: the case of detecting patient allergies in electronic health records. BMC Medical Informatics and Decision Making, 23(1), 188. https://doi.org/10.1186/s12911-023-02271-8
14. Khela, H., Khalil, J., Daxon, N., Neilson, Z., Shahrokhi, T., Chung, P., Wong, P. (2024). Real world challenges in maintaining data integrity in electronic health records in a cancer program. Technical Innovations Patient Support in Radiation Oncology, 29, 100233. https://doi.org/10.1016/j.tipsro.2023.10023
15. Chen, Y.P., Lo, Y.H., Lai, F., Huang, C.H. (2021). Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: Algorithm development and validation study. Journal of Medical Internet Research, 23(1), e25113. https://preprints.jmir.org/preprint/25113
16. Negro-Calduch, E., Azzopardi-Muscat, N., Krishnamurthy, R.S., Novillo-Ortiz, D. (2021). Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews. International Journal of Medical Informatics, 152, 104507. https://doi.org/10.1016/j.ijmedinf.2021.104507
17. Ramesh, J., Keeran, N., Sagahyroon, A., Aloul, F. (2021). Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning. In Healthcare, 9(11), 1450. https://doi.org/10.3390/healthcare9111450.
18. Miller, T.P., Getz, K.D., Li, Y., Demissei, B.G., Adamson, P.C., Alonzo, T.A., ... Aplenc, R. (2022). Rates of laboratory adverse events by course in paediatric leukaemia ascertained with automated electronic health record extraction: a retrospective cohort study from the Children's Oncology Group. The Lancet Haematology, 9(9), e678-e688.
19. Ros, F.M.N., Viejo, J.D.M. (2024). Preclinical evaluation of electronic health records (EHRs) to predict poor control of chronic respiratory diseases in primary care: A novel approach to focus our efforts. Journal of Clinical Medicine, 13(18), 5609. https://doi.org/10.3390/jcm1318560
20. Tang, A.S., Rankin, K.P., Cerono, G., Miramontes, S., Mills, H., Roger, J., ... Sirota, M. (2024). Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. Nature Aging, 4(3), 379-395. https://doi.org/10.1038/s43587-024-00573-8
21. Gandin, I., Saccani, S., Coser, A., Scagnetto, A., Cappelletto, C., Candido, R., ... Di Lenarda, A.

(2023). Deep-learning-based prognostic modeling for incident heart failure in patients with diabetes using electronic health records: A retrospective cohort study. Plos one, 18(2), e0281878. https://doi.org/10.1371/journal.pone.0281878

22. Adelusi, B.S., Osamika, D., Kelvin-Agwu, M.C., Mustapha, A.Y., Ikhalea, N. (2022). A deep learning approach to predicting diabetes mellitus using electronic health records. J Front Multidiscip Res, 3(1), 47-56.
23. Ruiz, V.M., Goldsmith, M.P., Shi, L., Simpao, A.F., Gálvez, J.A., Naim, M.Y., ... Tsui, F.R. (2022). Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records. The Journal of Thoracic and Cardiovascular Surgery, 164(1), 211-222. https://doi.org/10.1016/j.jtcvs.2021.10.060
24. Zhu, Y., Bi, D., Saunders, M., Ji, Y. (2023). Prediction of chronic kidney disease progression using recurrent neural network and electronic health records. Scientific Reports, 13(1), 22091. https://doi.org/10.1038/s41598-023-49271-2
25. Mantey, E.A., Zhou, C., Srividhya, S.R., Jain, S.K., Sundaravadivazhagan, B. (2022). Integrated blockchain-deep learning approach for analyzing the electronic health records recommender system. Frontiers in Public Health, 10, 905265. https://doi.org/10.3389/fpubh.2022.905265
26. Akter, S., Liu, Z., Simoes, E.J., Rao, P. (2025). Using machine learning and electronic health record (EHR) data for the early prediction of Alzheimer's Disease and Related Dementias. The Journal of Prevention of Alzheimer's Disease, 100169. https://doi.org/10.1016/j.tjpad.2025.100169
27. Singh, V., Asari, V.K., Rajasekaran, R. (2022). A deep neural network for early detection and prediction of chronic kidney disease. Diagnostics, 12(1), 116. https://doi.org/10.3390/diagnostics12010116
28. Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., ... Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. IEEE access, 9, 17312-17334. https://doi.org/10.1109/ACCESS.2021.3053763
29. Debal, D.A., Sitote, T.M. (2022). Chronic kidney disease prediction using machine learning techniques. Journal of Big Data, 9(1), 109. https://doi.org/10.1186/s40537-022-00657-5
30. Mondol, C., Shamrat, F.J.M., Hasan, M.R., Alam, S., Ghosh, P., Tasnim, Z., ... Ibrahim, S.M. (2022). Early prediction of chronic kidney disease: A comprehensive performance analysis of deep learning models. Algorithms, 15(9), 308. https://doi.org/10.3390/a15090308