Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Generative AI–Driven Clinical Decision Support Framework Using Large Language Models

Jahnavi Anilkumar Kachhia
Independent Researcher, California State University Fullerton
Corresponding Author: Jahnavi Anilkumar Kachhia **E-mail**: jahnavikachhia2025@gmail.com

| ABSTRACT

Early disease detection aids in the correct diagnosis and treatment of illnesses. A Clinical Decision Support System (CDS) helps identify illnesses and choose the best course of therapy. This paper presents a Generative AI-powered Clinical Decision Support Architecture based on Large Language Models (LLMs) to predict diseases and support diagnoses. The suggested architecture incorporates both structured clinical information and high-level preprocessing, feature selection, and class-balancing algorithms to increase the predictive accuracy. Experiments were conducted on 400 patient records from the UCI Chronic Kidney Disease (CKD) dataset. GPT-4o was used to learn more complex clinical patterns and aid in diagnostic decision-making. The recommended framework performed well, as evidenced by the accuracy of 99.17, sensitivity of 99.98, specificity of 98.70, F1-score of 98.85, Matthews Correlation Coefficient (MCC) of 98.21, and AUROC of 0.996. These findings are far more effective than conventional ML models and currently available LLM-based clinical methods. The high sensitivity yields a low rate of false negatives, which is essential in the early detection of disease, whereas the high specificity lowers the wrong diagnosis of healthy patients. Altogether, the suggested generative AI-based solution is powerful, consistent, and effective in clinical contexts, which underscores the potential of large language models (LLM) in medical decision support systems of the next generation.

| KEYWORDS

Clinical Decision Support System, Chronic Kidney Disease, Electronic Health Records, Machine Learning, Disease Prediction, Large Language Models

| ARTICLE INFORMATION

ACCEPTED: 01 December 2024 **PUBLISHED:** 25 December 2024 **DOI:** 10.32996/jcsts.2024.6.5.31

I. Introduction

The widespread digitization of patient information brought about by the expanding usage of Electronic Health information (EHRs) has completely transformed healthcare systems [1], [2]. With such data, which is cleansed and preprocessed systematically, there a solid basis of advanced analytics and clinical decision support (CDS). Nevertheless, with the increasing volume, heterogeneity and longitudinal character of clinical data, the complexity of medical decision-making has significantly increased. Traditional rule-based approaches are no longer applicable, as physicians must combine laboratory findings, comorbidities, treatment history, and other patient-specific risk factors. The capacity of ML-based Clinical Decision Support (CDS) systems to recognize complex, non-linear connections and work with high-dimensional data, on the other hand, has made them more and more popular [3]. In this context, chronic kidney disease (CKD) is an important application domain of intelligent decision support because it is very prevalent in the global population, and progressive. CKD disturbs the normal functions of the kidneys and causes severe systemic complications, such as cardiovascular disease, Autism Spectrum [4], bone diseases, and neurological disorders . It is still on the increase in the world, and most of these conditions have been fueled by risk factors like high blood pressure, diabetes, and cardiovascular diseases.

Thus, reducing mortality, shortening the course of the disease, and improving patients' quality of life all depend on early detection and treatment. In the diagnosis and monitoring of CKD, conventional diagnostic indicators, including glomerular filtration rate (GFR) and urine albumin levels, remain crucial. Nonetheless, such clinical indication has drawbacks in disease progression and the fact that physiological intricacy of CKD is complex to indicate [5]. This has led to the application of machine learning (ML) methods, which have proven to have great potential in deriving useful patterns of complex clinical data and enhancing the accuracy of diagnosis [6]. In spite of these developments, most traditional ML models are black box models, which restrict their interpretability and clinical trust. Recent trends in generative artificial intelligence (AI) and large language models (LLMs) provide an opportunity to replace it with a promising alternative with high predictive power and more advanced reasoning and representation learning. This paper is inspired by these developments to introduce a Generative AI-based Clinical Decision Support Framework with Large Language Models to provide accurate, robust, and clinically reliable assistance in the CKD diagnosis and overcome the major drawbacks of the traditional ML methods. The following are the contributions of the paper are:

- Establishes the feasibility of adapting GPT-4o beyond natural language tasks to high-accuracy clinical diagnosis.
- Demonstrates strong diagnostic reliability through balanced sensitivity, specificity, F1-score, and MCC, even under class imbalance.
- Provides empirical evidence that generative AI can outperform traditional machine learning and existing LLM-based models in clinical decision support.
- Validates the proposed framework through comprehensive experimental analysis and comparison with state-of-the-art methods.
- Minimizes false negatives, making the framework highly suitable for early disease detection and clinical screening.

Data imbalance, missing values and inadequate reliability of early and precise clinical decision-making is also difficult due to existing models. This work is motivated by the need for a more reliable and clinically sensitive decision support system. The novelty is that by using a large language model based on generative models, it is possible to obtain highly balanced and near-perfect diagnostic performance. The suggested solution proves more sensitive, specific, and generally healthier, establishing a new performance standard and outlining the utility of generative AI in practical clinical decision support.

A. Paper structure

The paper structure is as follows: Section II presents the previous study on clinical decision support. The proposed methodology is presented in Section III, and the results and comparison are presented in Section IV. Conclusion with limitations and future work in Section V.

II. Literature Review

Existing literature has covered both traditional and DL frameworks in clinical decision support systems.

Periya Nayaki *et al*. (2024) present the development of a Clinical Decision Support System (CDSS) utilizing DL techniques, specifically focusing on RNNs and LSTM networks. Among the models, the Bidirectional LSTM network outperformed traditional machine learning methods, demonstrating its more advanced ability to handle sequential clinical data has an accuracy of 90.16% [7].

Raza and Ding (2024) identify and rank the items most pertinent to healthcare practitioners using a DL model. The foundation of both the ranker and the retriever is a pipeline of pre-trained transformer models. The model's performance was evaluated using a range of metrics and compared with comparable baseline models. In comparison to the second-best performing baseline, the findings show that the suggested model generates a performance improvement of about 12.3% macro-average F1 [8].

Kim *et al*. (2023) used validation datasets to compare DeepSEPS with other conventional early warning score systems. In terms of AUROC, DeepSEPS outperformed SOFA, achieving 0.7888 and 0.8494 for sepsis and septic shock, respectively. Additionally, DeepSEPS had the highest AUROC (0.9346) when septic shock and sepsis were developing [9].

Kim *et al.* (2022) A DL model for CDSSs that can be used to supervised learn three classes of sleep phases using just single-channel EEG data (C4-M1) may be created by combining CNN with a transformer. Comparable to human experts, the generated model produced an overall accuracy of 91.4%. 94.3%, 91.9%, 91.9%, and 90.6%, respectively, were the model's accuracy in normal, mild, moderate, and severe instances of obstructive sleep apnoea [10].

Shahzad *et al.* (2021) suggest a framework for predicting health problems based on LSTM in order to correct noisy and unbalanced data and convert it into a format that can be used to provide precise predictions. The proposed model is capable of efficiently training, validating, and test noisy data by overcoming the lack of training data using transfer learning and consistently achieving outcomes of around 90% better than the state-of-the-art ML and DL approaches [11].

Kormilitzin *et al*. (2020) propose a named-entity recognition (NER) paradigm for clinical NLP. The model recognises seven categories: medication names, dosage, frequency, strength, form, duration, and mode of administration. In all seven categories, the model's micro-averaged F1 score was 0.957. The trained NER model performed badly (F1 = 0.762) when applied directly to CRIS data, but it performed very well (F1 = 0.944) after being fine-tuned on a small sample of CRIS data [12].

Geng *et al*. (2020) Analyse NLP and electronic medical records (EMRs)-based model-based reasoning (MBR) clinical diagnostic algorithms in integrated medicine. Using symptom patterns, the Word2Vec CNN MBR algorithms demonstrated exceptional effectiveness in diagnosing lung ailments (accuracy of 0.9586 on the test dataset). Additionally, the Word2Vec CNN MBR and RBR combination performed quite well (accuracy of 0.9229 in the test data set) [13].

A. Research Gap

The majority of the existing CDSS research presents a good outcome, but is still limited. All models are effective for a particular disease, dataset, or task, but they have limitations in their generalizability across different clinical environments. The management of multimodal data, noisy records and imbalanced records are only partially considered as well as the privacy concerns. The combination of sequential models, transformers, and domain adaptation into coherent structures is very rare. Explainability and real-world deployment also receive little attention. The disconnect is the creation of scalable, interpretable, and transferable CDSS that is capable of integrating diverse data sources and acts reliably across diverse clinical areas.

III. Methodology

This paper uses Chronic Kidney Disease (CKD) data and preprocesses with KNN and mode imputation to fill in blank data and categorical encoding, along with minmax scaling. The Chi-squared feature selection, an 80:20 train test split, and a SMOTE were used to balance the classes. An evaluation of performance was conducted using accuracy, sensitivity, specificity, F1-score, MCC, and AUROC with a GPT-4o-based model. Fig. 1 present the CDS framework.

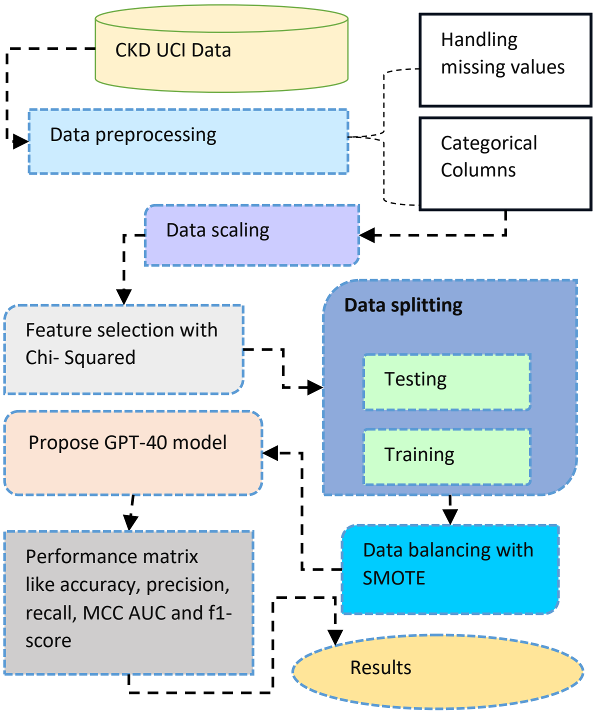

Fig. 1. Clinical Decision Support SystemFramework

All the steps of proposed framework are explored in next section.

A. Data Gathering

The dataset used in this work includes CKD patient records from 400 patients from the UC Irvine (UCI) Machine Learning repository. It has 25 qualities total, of which 1 is a dependent variable (class), and 24 are independent variables (features). The total number of records is 400, of which 250 are from 150 people who do not have CKD, and 150 are from patients who do. Out of the twenty-four attributes, eleven are numerical, and thirteen are not. The numbers are obtained from blood and urine tests performed on the patients. Either 1 or 0 class values are present in the output columns. CKD is present in the patient if the result is 1, but not if it is 0.

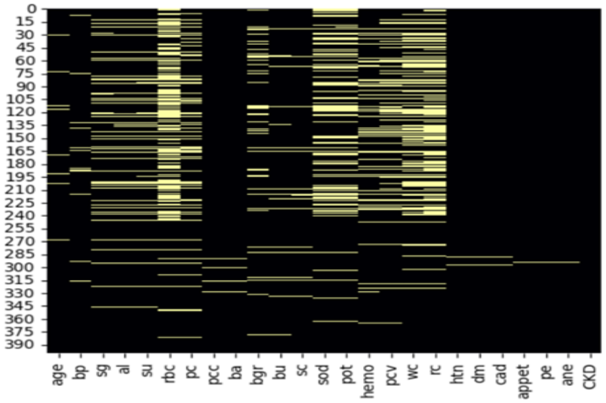

Fig. 2. Heatmap for Missing values in data

Fig. 2 shows a heatmap of missing data and various medical characteristics, with yellow circles indicating missing data and black circles denoting provided data. Variables like red blood cells (rbc), pus cells (pc), packed cell volume (pcv), white blood cell count (wc), and red blood cell count (rc) have significant gaps whereas aspects like age, blood pressure and chronic disease indicators (htn, dm, cad) are more complete. This visualization shows that there is an uneven distribution of missing information that requires meticulous preprocessing approaches that could either be imputation or feature dropping to guarantee quality model development.

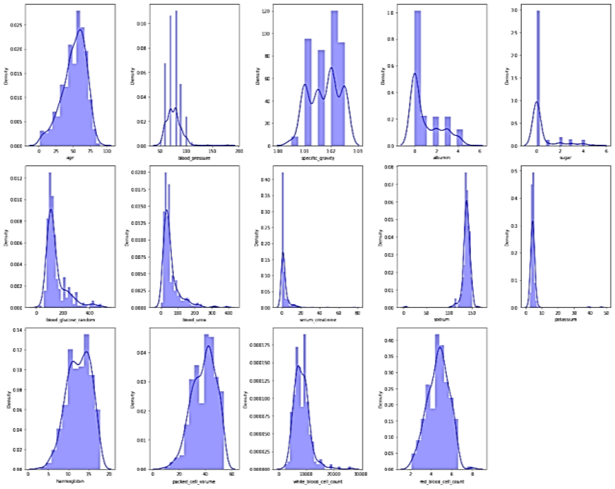

Fig. 3. Plot for clinical features Distributions

Fig. 3 presents the patterns of the major clinical characteristics in the data includes, among other things, age, blood pressure, specific gravity, albumin, blood sugar, blood glucose, blood urea, serum creatinine, serum potassium, haemoglobin, packed cell volume, white blood cell count, and red blood cell count. A majority of the variables which include age, hemoglobin, and packed cell volume have more or less normal distribution, but others, including blood urea, serum creatinine, and albumin, are highly skewed, which shows that it contains outliers or extreme values in the clinical data. Such differences indicate heterogeneity in clinical measurements, and feature-specific preprocessing and normalization are essential for successful model training and CKD prediction.

B. Data Pre-processing and Data Cleaning

The obtained data has to be organised or processed in order to be analysed. Missing values are present in the dataset the data must be pre-processed before being fed into the models. Simply eliminating records with missing values from the dataset is not appropriate due to the small size of the chosen dataset; nevertheless, as Fig. 2 illustrates, it has a sizable portion of missing data. Consequently, Use mode imputation for categorical data and KNN imputation for numerical values.

C. Categorical columns

The dataset includes numeric columns for the category variables. The values "good," "present," "yes," and "normal" are mapped to 1, whereas the values "bad," "not present," "no," and "abnormal" are mapped to 0. The dataset was subsequently processed, and all of the values were normalised using scaling.

D. Data Scaling

A crucial component of preprocessing is data scaling, which modifies the range of feature values while preserving the data's inherent integrity. Min–max scaling is a popular data scaling method that rescales data values to lie within a specified range, typically 0 to 1. Here is Formula 1 for min-max scaling:

$$Min - Max\ Scaling(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

The feature values in the research were normalized using min–max scaling to enhance and streamline data processing.

E. Feature Selection

The relationship between independent traits and the target class is identified using the chi-squared feature selection approach. Because it can handle categorical data with robustness and doesn't assume anything, the chi-squared test was employed in this study to identify features based on the data distribution. This test also calculated the chi-squared value for each independent characteristic and the target class. The chi-squared scores of the top nine characteristics as determined by the chi-squared test.

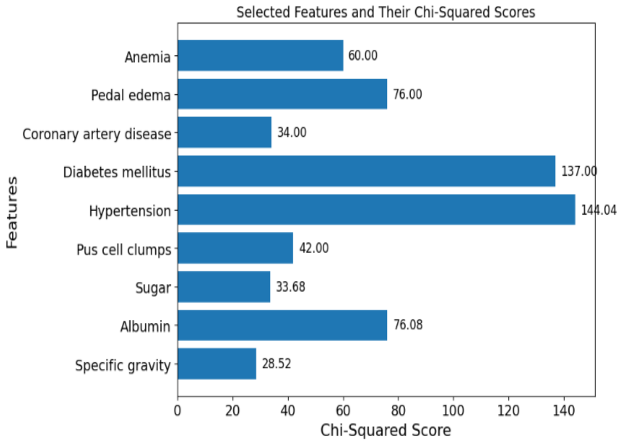

Fig. 4. Top 9 Chi-Squared Feature Importance Scores

Fig. 4 shows the nine most important medical features according to the Chi-Squared values, which are the values of the medical features in predictive model. The strongest contributors are hypertension (144.04) and diabetes mellitus (137.00), then albumin (76.08) and pedal edema (76.00). The moderate importance is found in anemia and pus cell clumps whereas coronary artery disease, sugar, and specific gravity have less but significant impact. This distribution is based on the fact that chronic conditions and key laboratory indicators are the most powerful determinants of model performance.

F. Train-Test Split

The dataset has to be divided into training and testing when data processing was finished. The remaining 80% of the dataset is used for training, with 20% reserved for testing.

G. SMOTE for Balancing

The SMOTE strategy, which enhances the illustration of the minority class, mitigates the effects of class imbalance. By adding synthetic examples, SMOTE improves learning by giving the classifier a more evenly distributed training set. Two strategies to address class imbalance were to identify the minority class and determine the imbalance ratio. Next, the SMOTE method was used to create synthetic samples by interpolating between minority examples that were selected at random and their closest neighbours. To provide a more representative dataset for model training, these artificial examples were added to the training set until the appropriate balance was reached.

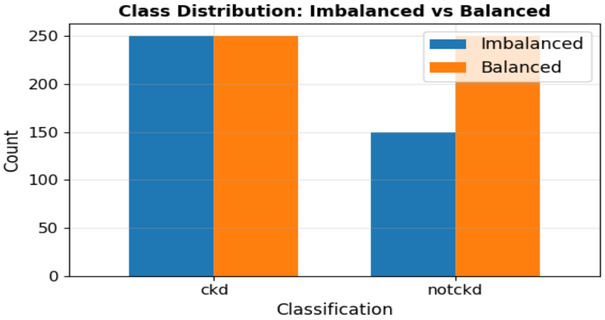

Fig. 5. Bar Graph for Imbalance and Balance Class Distribution

Fig. 5 presents a bar graph of the distribution of imbalanced and balanced classes in CKD prediction. The class of ckd has 250 samples, whereas the class of not ckd has only 150 samples in the imbalanced dataset, showing the skewness in representation. Once balancing is complete, the two classes are set to 250 samples per category to represent both categories equally. The visualization clearly shows how balancing helps correct class imbalance, making it more appropriate for creating reliable classification models.

H. Proposed GPT-4o Model

The feedforward network is position-wise fully coupled and consists of a multi-head self-attention mechanism each of the two main sub-layers that make up GPT-4's multi-layer design. For tasks like translation, text production, and summarisation that call for context comprehension, the self-attention mechanism allows the model to determine the relative relevance of different tokens in the input sequence [14]. Furthermore, GPT-4 incorporates positional encodings to account for the order of the word sequence, resolving a major issue with the original Transformer model—namely, its inability to automatically capture sequential data.

The pre-training process is based on a large, diverse set of text and multimodal data and is run with substantial parallelization on large-scale GPUs/TPUs to train billions of parameters. This enables GPT-4o to capture long-range dependencies, semantic relationships, and syntactic patterns, leading to excellent performance on tasks such as question answering, reasoning, summarising, and conversational discourse. The GPT-4o goes through supervised fine-tuning after pre-training and Reinforcement Learning from Human Feedback (RLHF), during which human evaluators check outputs for accuracy, safety, and relevance. This stage of fine-tuning makes the model more in line with what people want, reduce biases and improve reliability, which is why this model can be used in sensitive settings, such as clinical decision support, where accuracy, interpretability, and safety are of utmost importance. Additional refinement of the system based on medical literature and clinical guidelines to domain-specific fine-tuning makes it more useful in the healthcare environment.

I. Evaluation Parameters

The classifiers' ability to identify diabetes was evaluated using metrics such as AUROC, precision, specificity, sensitivity, and accuracy. This measurement is computed using a confusion matrix, which performs a matrix-like comparison between the actual and predicted classes. Therefore, the following estimated values are provided:

- **True positive (TP):** It estimates the proportion of the expected class's positive events when the actual class was equally positive.
- **True negative (TN):** It estimates the probability that the expected class would have negative outcomes, and the real class did as well.
- **False positive (FP):** It analyses the frequency with which the projection class was positive, but the actual class was negative.
- **False negative (FN):** It examines how frequently the actual class was positive while the projected class was negative.

Then, several assessment metrics—which are as follows—are altered:

1) ***Accuracy:*** The accuracy is calculated by dividing the total number of accurate forecasts by the total number of predictions. It is computed using equation (2):

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+Fp+TN+FN}} \quad (2)$$

2) ***Sensitivity:*** It is the proportion of positives among all positives that are accurately detected. It is computed using equation (3):

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

3) ***Specificity:*** It is defined as the percentage of all negatives that are accurately detected. It is computed using equation (4):

$$Specificity = \frac{TN}{TN+FP} \qquad (4)$$

4) ***F1-score:*** It is a measure of the harmony between sensitivity and precision on a single parameter, and its values range from 0 to 1. The closer it is to 1, the better. Equation (5) determines it:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (5)$$

5) ***MCC:*** MCC is an efficient statistic for assessing performance on unbalanced datasets since it takes into account every component of the confusion matrix. Equation (6) calculates it:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (6)$$

6) ***AUROC:*** It is an assessment metric that manipulates the true positive and false positive rates, respectively, to create a result. This number is closest to 1, which is what a good model considers.

IV. Results and Discussion

The AMD Ryzen 5 5600X 6-core CPU, 32 GB of RAM, a GeForce 3060 Ti (NVIDIA), Python 3.7.11, and PyTorch 1.11.0 made up the experimental setup. To implement the metamodel, Python was selected as the programming language, and the scikit-learn package was used. The metamodel was trained in Google Colab. As shown in Table I, five assessment metrics were used to evaluate LLM's performance: MCC, F1-score, recall, accuracy, and precision. The model has 99.17% accuracy, indicating strong overall predictive power. The sensitivity of 99.98% indicates its usefulness in supporting clinical decision-making by accurately determining the cases of CKD. Also, the high F1-score (98.85) and MCC (98.21) indicate that GPT-4o is appropriate for CKD diagnosis despite possible class imbalance.

TABLE I. EXPERIMENT OUTCOME OF LLM MODEL ON CKD DATA

Matrix	**GPT-4o**
Accuracy	99.17
Sensitivity	99.98
Specificity	98.70
F1-score	98.85
MCC	98.21

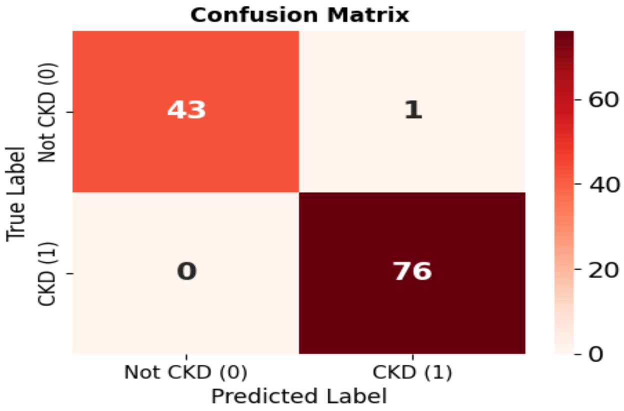

Fig. 6. Confusion Matrix of GPT 40 model

Fig. 6 shows the confusion matrix of GPT-40 model on predicting CKD in which the true labels are compared to the predicted labels. The table indicates that 43 non-CKD patients were appropriately categorized, with only one patient being wrongly categorized as CKD, and all 76 CKD patients were correctly identified without any false negatives. This distribution shows that the model has strong diagnostic potential, with very high accuracy and reliability, and a low error rate, which may be utilised to distinguish between patients with and without CKD.

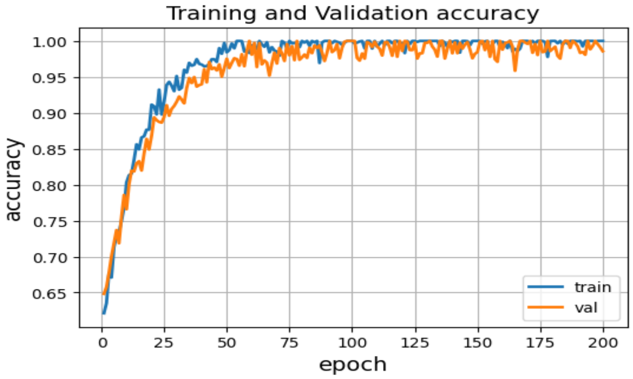

Fig. 7. Plot Accuracy Curve of GPT-4o Model

The GPT-4o model's training and validation accuracy curve with 200 epochs is displayed in Fig. 7. The accuracy is high at the first training stage, which means that the features are learned well and, however, in both training and validation sets, it progressively gets closer to 100%. The tight correspondence between the two curves with a small difference shows that there is stable learning behavior, high generalization performance and very small overfitting during the training process.

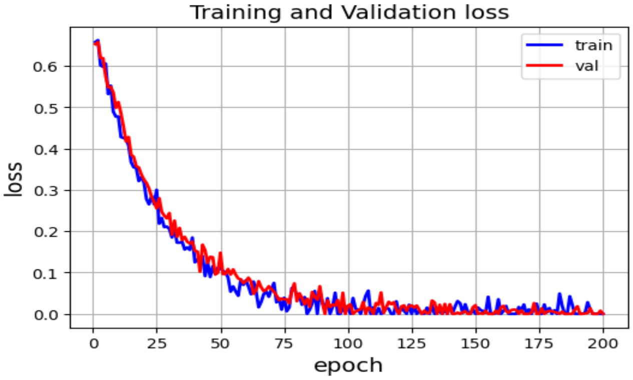

Fig. 8. ROC Curve of GPT-40 Model

The ROC curve for the GPT-40 model is shown in Fig. 9 and plots TPR (sensitivity) versus FPR (1-specificity). The blue curve shows the model's discriminative power compared with the red dashed diagonal line, which represents random classification. An orange mark at the upper-left corner shows that there is a critical threshold that has a high sensitivity yet low false positive rate. The value is 0.996, indicating that the GPT-40 model is highly effective in its classification performance, as it can effectively distinguish between CKD and non-CKD patients.

TABLE II. PERFORMANCE COMPARISON OF EXISTING AND PROPOSE MODEL ON CLINICAL AND CKD DATASETS

Classifiers	**Datasets**	**Accuracy**	**Specificity**	**Sensitivity**	**F1-score**	**AUC**
ChatGPT ADA: RF [15]	clinical-trial dataset	0.892	0.903	0.884	0.894	0.954
GatorTronGPT-20B [16]	82 billion clinical texts	-	0.476	0.521	0.500	-
PubMedBERT [17]	Acute-renal-failure-from-clinical-notes	0.832	0.952	0.200	0.276	0.596
CXR model [18]	MIMIC-IV	0.81	0.23	0.86	0.36	0.71
MLP [19]		0.64	0.63	0.75	-	0.74
XGBoost [20]		78.0	27.3	96.7	-	-
Random forest [21]		0.655	-	-	0.442	0.800
GRU-D [22]	MIMIC-III	94.0	-	-	43.1	89.1
SVM [23]		0.76	0.75	0.78	-	0.86
RoBERTa-OHFT-DeCLUTR [24]		0.557	0.565	0.557	0.539	0.831

KNN [25]	**CKD dataset**	0.74	0.79	0.78	0.78	-
Ridge [26]		95	89.47	97.14	93.15	-
ANN [27]		96.25	94	90.32	97	-
GB [28]		0.662	0.674	0.69	0.658	-
Propose		**99.17**	**98.70**	**99.98**	**98.85**	**0.996**

Table II shows a quantitative comparison of current models and the proposed GPT-4o framework on clinical and CKD data. Current methods like ChatGPT ADA: RF report the accuracy of 89.2% whereas PubMedBERT has the accuracy of 83.2% which restricts its applicability to clinical use. CXR and GRU-D models are characterized by high sensitivity (86.0% and 94.0%, respectively) but low specificity and F1 Scores, indicating that predictions are unbalanced. The conventional machine learning classifiers, MLP (64.0% accuracy), SVM (76.0%), and KNN (74.0%), have moderate and haphazard performance. Ridge (95.0 percent accuracy) and ANN (96.25 percent accuracy) still do better on the CKD dataset than a number of baselines but is still worse than the proposed model. Conversely, the suggested GPT-4o has the best overall performance which outperforms existing methods in terms of superior balance, robustness and diagnostic reliability.

The proposed GPT-4o framework offers important benefits over existing frameworks, achieving higher accuracy, sensitivity, and specificity, and a higher AUROC, providing reliable, balanced CKD diagnosis. It has almost ideal sensitivity, which guarantees the fewest false negatives, which is important in clinical screening, and high specificity, which minimizes false alarms. Consistent high F1-score and MCC demonstrate efficacy even with class imbalance, illustrating good generalization and the appropriateness of GPT-4o for real-world clinical decision support systems.

V. Conclusion And Future Work

Generative artificial intelligence has become a paradigm for improving clinical decision-making through more accurate, data-driven insights. This paper proposes a clinical decision support model trained with a large language model to predict diseases. The experimental outcome on the CKD data proves that the proposed GPT-4o framework provides superior performance with obtaining 99.17 percent accuracy, MCC of 98.21, and an automatic under-recurve (AUROC) of 0.996. These results show that the predictive accuracy is very good but also high robustness and balanced classification which are necessary to have clinical reliability. Compared to the existing machine learning, deep learning, and LLM-based systems, the proposed model is always better than the previous systems such as ANN, Ridge, SVM, and other clinical language models, which report relatively low accuracy, sensitivity, or generalization ability. The low rate of false-negatives also emphasizes its appropriateness in clinical screening and early illness detection. Overall, the results support the hypothesis that generative large language models can effectively facilitate structured clinical diagnosis and are a promising future for clinical decision support systems.

A. Recommendations

The suggested generative AI-based clinical decision support system can be successfully implemented in hospital information systems to enable clinicians to screen for and diagnose CKD in the early stages. Such systems are intended to be implemented as decision-support systems, not to replace clinical experience, as human control is required in critical situations. Data governance, validation of the model and frequent updates with new clinical data are all required to ensure reliability, safety and trust in real-world healthcare environment.

B. Challenges

Although the proposed generative AI-driven framework performs well, there are still several challenges. The reason is that the model's current validation is based on a relatively small, single-source dataset and may not be applicable to other clinical groups. There is also the challenge of relying on quality and well-documented data on clinical data since the real-life medical records usually have noises and discrepancies. Besides that, large language models have high computational and resource costs, which limits their application in low resource health care. Lastly, there are problems concerning the model's interpretability, clinical trust, data privacy, and regulatory compliance, which should be taken into consideration before real-world application.

C. Future Work

Further studies can build on this framework to make predictions of multiple diseases based on bigger and more varied clinical data in actual hospital environments. Using longitudinal patient records and unstructured clinical notes could also improve diagnostic accuracy. Also, enhancing the explainability and interpretability of generative AI decisions will aid in clinical adoption and external validation and real-time deployment studies will aid in evaluating scalability and generalizability to other healthcare systems.

References

[1] C. Tayal, "Big Data Pipeline Optimisation for Electronic Health Records (EHR)," *Int. J. Artif. Intell. Data Sci. Mach. Learn.*, vol. 5, no. 3, pp. 121–127, 2024, doi: 10.63282/3050-9262.ijaidsml-v5i3p113.

[2] Y. Macha, "A Review of Cloud-Based CRM Systems in Healthcare: Advances , Tools , Challenges , and Best Practices," *Int. J. Curr. Eng. Technol.*, vol. 12, no. 6, pp. 848–855, 2022.

[3] V. Pal, "Foundation Models for Multi-Modal Clinical Decision Support Systems," *ESP J. Eng. Technol. Adv.*, vol. 2, no. 2, pp. 183–191, 2022.

[4] S. Thangavel, "A System And Method For Early Detection Of Autism Spectrum Disorder Using Machine Learning," 202411064771, 2024

[5] C. Tayal, "AI-Enhanced ETL Framework for Improving Data Quality in Clinical Decision Support Systems," *Int. J. Artif. Intell. Data Sci. Mach. Learn.*, vol. 5, no. 2, pp. 116–120, 2024.

[6] D. Saif, A. M. Sarhan, and N. M. Elshennawy, "Deep-kidney: an effective deep learning framework for chronic kidney disease prediction," *Heal. Inf. Sci. Syst.*, vol. 12, no. 3, pp. 01–22, 2024, doi: 10.1007/s13755-023-00261-8.

[7] A. Periya Nayaki, M. S. Arrchit Ramana, S. D. Prithivi Raj, M. S. Thanabal, and N. K. Jeyakumar, "Clinical Decision Support System Using Recurrent Neural Network," in *2024 1st International Conference on Data, Computation and Communication, ICDCC 2024*, 2024. doi: 10.1109/ICDCC62744.2024.10961004.

[8] S. Raza and C. Ding, "Improving Clinical Decision Making With a Two-Stage Recommender System," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 21, no. 5, pp. 1180–1190, Sep. 2024, doi: 10.1109/TCBB.2023.3318209.

[9] T. Kim *et al.*, "Development and Validation of Deep-Learning-Based Sepsis and Septic Shock Early Prediction System (DeepSEPS) Using Real-World ICU Data," *J. Clin. Med.*, 2023, doi: 10.3390/jcm12227156.

[10] D. Kim, J. Lee, Y. Woo, J. Jeong, C. Kim, and D. K. Kim, "Deep Learning Application to Clinical Decision Support System in Sleep Stage Classification," *J. Pers. Med.*, 2022, doi: 10.3390/jpm12020136.

[11] Y. Shahzad, H. Javed, H. Farman, J. Ahmad, B. Jan, and A. A. Nassani, "Optimized Predictive Framework for Healthcare Through Deep Learning," *Comput. Mater. Contin.*, vol. 67, no. 2, pp. 2463–2480, 2021, doi: 10.32604/cmc.2021.014904.

[12] A. Kormilitzin, N. Vaci, Q. Liu, and A. Nevado-Holgado, "Med7: a transferable clinical natural language processing model for electronic health records," *Artif. Intell. Med.*, Apr. 2020, doi: 10.1016/j.artmed.2021.102086.

[13] W. Geng *et al.*, "Model-based reasoning of clinical diagnosis in integrative medicine: Real-world methodological study of electronic medical records and natural language processing methods," *JMIR Med. Informatics*, 2020, doi: 10.2196/23082.

[14] A. Radfort, K. Narasimhan, T. Salimans, and I. Sutskever, "(OpenAI Transformer): Improving Language Understanding by Generative Pre-Training," *OpenAI*, 2018.

[15] S. Tayebi Arasteh *et al.*, "Large language models streamline automated machine learning for clinical studies," *Nat. Commun.*, vol. 15, no. 1, p. 1603, Feb. 2024, doi: 10.1038/s41467-024-45879-8.

[16] C. Peng *et al.*, "A study of generative large language model for medical research and healthcare," *npj Digit. Med.*, vol. 6, no. 1, p. 210, Nov. 2023, doi: 10.1038/s41746-023-00958-w.

[17] O. Litake, B. H. Park, J. L. Tully, and R. A. Gabriel, "Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes," *J. Am. Med. Informatics Assoc.*, vol. 31, no. 6, pp. 1404–1410, May 2024, doi: 10.1093/jamia/ocae081.

[18] J. Lin *et al.*, "Development and Validation of Multimodal Models to Predict the 30-Day Mortality of ICU Patients Based on Clinical Parameters and Chest X-Rays," *J. Imaging Informatics Med.*, vol. 37, no. 4, pp. 1312–1322, Mar. 2024, doi: 10.1007/s10278-024-01066-1.

[19] X. Zhang, N. Fei, X. Zhang, Q. Wang, and Z. Fang, "Machine Learning Prediction Models for Postoperative Stroke in Elderly Patients: Analyses of the MIMIC Database," *Front. Aging Neurosci.*, vol. 14, Jul. 2022, doi: 10.3389/fnagi.2022.897611.

[20] J. Lin, C. Gu, Z. Sun, S. Zhang, and S. Nie, "Machine learning-based model for predicting the occurrence and mortality of nonpulmonary sepsis-associated ARDS," *Sci. Rep.*, vol. 14, no. 1, p. 28240, Nov. 2024, doi: 10.1038/s41598-024-79899-7.

[21] L. Hempel, S. Sadeghi, and T. Kirsten, "Prediction of Intensive Care Unit Length of Stay in the MIMIC-IV Dataset," *Appl. Sci.*, 2023, doi: 10.3390/app13126930.

[22] S. Wang, M. B. A. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann, "MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, New York, NY, USA: ACM, Apr. 2020, pp. 222–235. doi: 10.1145/3368555.3384469.

[23] X. Jiang, W. Dai, and Y. Cai, "Comparison of machine learning algorithms to SAPS II in predicting in-hospital mortality of fractures of the pelvis and acetabulum: analyzes based on MIMIC-III database," *All Life*, vol. 15, no. 1, pp. 1000–1012, Dec. 2022, doi: 10.1080/26895293.2022.2125448.

[24] N. Taylor, D. Schofield, A. Kormilitzin, D. W. Joyce, and A. Nevado-Holgado, "Developing healthcare language model embedding spaces," *Artif. Intell. Med.*, vol. 158, p. 103009, Dec. 2024, doi: 10.1016/j.artmed.2024.103009.

[25] C. Kaur, M. S. Kumar, A. Anjum, M. B. Binda, M. R. Mallu, and M. S. Al Ansari, "Chronic Kidney Disease Prediction Using Machine Learning," *J. Adv. Inf. Technol.*, 2023, doi: 10.12720/jait.14.2.384-391.

[26] M. S. Arif, A. U. Rehman, and D. Asif, "Explainable Machine Learning Model for Chronic Kidney Disease Prediction," *Algorithms*, vol. 17, no. 10, p. 443, Oct. 2024, doi: 10.3390/a17100443.

[27] C. Mondol *et al.*, "Early Prediction of Chronic Kidney Disease: A Comprehensive Performance Analysis of Deep Learning Models," *Algorithms*, vol. 15, no. 9, p. 308, Aug. 2022, doi: 10.3390/a15090308.

[28] S. G, S. J, C. K, and V. K. V, "Prediction Of Chronic Kidney Disease Using Machine Learning," *IJARCCE*, vol. 13, no. 4, Apr. 2024, doi: 10.17148/IJARCCE.2024.134160.