

AI-Driven CKD Diagnosis: A Full-Stack Approach to Early Prediction

¹Y. Priya, ²P. Sreeja, ³M. Nandini, ⁴V. Ravi Shankar, ⁵B. Sahith, ⁶M. Goutham

¹Department of CSE- DATA SCIENCE

Sphoorthy Engineering College, Hyderabad, India

Abstract: *Chronic Kidney Disease (CKD) is a major global health issue, primarily caused by underlying conditions such as diabetes and hypertension. Early prediction of CKD is essential to prevent its progression to end-stage renal disease (ESRD), which often results in severe health complications and a significant reduction in quality of life. Traditional diagnostic methods for CKD are often invasive, expensive, and not easily accessible in low-resource settings, making early detection a persistent and difficult challenge. This project proposes an innovative, non-invasive, and cost-effective approach for early CKD prediction using an unsupervised machine learning framework. The system leverages a combination of algorithms, including K-Means, DBSCAN, Isolation Forest, and Autoencoders, along with feature selection and dimensionality reduction techniques, to effectively classify patients into CKD and non-CKD categories. Unlike supervised methods, this framework does not require labeled training data, making it ideal for real-world applications where such data may be scarce or unavailable. Clinical data used to train and validate the model demonstrated that the integration of these unsupervised techniques could achieve a classification accuracy between 85% and 90%. This level of accuracy highlights the strong potential of the system to support medical professionals in identifying CKD at an early stage, even in areas where specialized healthcare services are limited or nonexistent. By providing timely insights into patient risk profiles, the model facilitates earlier intervention and supports more personalized treatment strategies. Ultimately, this project aims to contribute to a more accessible, accurate, and efficient healthcare system, offering a scalable solution that can revolutionize CKD screening and improve patient outcomes worldwide.*

Keywords: Chronic Kidney Disease (CKD), Early prediction, Unsupervised Learning, K-Means Clustering, Autoencoders, Isolation Forest, Non-Invasive Diagnosis, Clinical Data Analysis

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a progressive condition characterized by a gradual loss of kidney function over time. It is a major public health issue globally, often caused by underlying conditions such as diabetes and hypertension. Early prediction of CKD is critical to preventing severe outcomes like end-stage renal disease (ESRD), which requires dialysis or kidney transplantation. However, current diagnostic methods are often invasive, expensive, and may not be readily available in under-resourced regions.

This project proposes an innovative, non-invasive approach for the early prediction of CKD using unsupervised machine learning techniques. By applying algorithms such as K-Means, DBSCAN, Isolation Forest, and Autoencoders, the system analyzes clinical data to differentiate between CKD and non-CKD cases. Through effective feature selection and

dimensionality reduction, the framework achieves high classification accuracy, enabling early diagnosis without the need for labeled data. This approach aims to assist healthcare providers in making timely decisions, especially in areas with limited access to specialized medical care.



II. RELATED WORKS

Several studies have explored machine learning techniques for the early detection of Chronic Kidney Disease (CKD). Traditionally, supervised learning models such as Support Vector Machines (SVM), Decision Trees, Random Forests, and Logistic Regression have been widely applied to classify CKD based on clinical datasets. These models have demonstrated high accuracy but depend heavily on the availability of labeled data, which can be limited or unreliable in many real-world healthcare settings.

In contrast, unsupervised learning methods offer an alternative that does not require labeled datasets. Some studies have investigated clustering techniques like K-Means or hierarchical clustering to identify patterns in unlabeled clinical data, but these often lack robustness when dealing with high-dimensional or noisy data. Autoencoders and Isolation Forests have also been explored for anomaly detection in medical datasets, offering promising results for detecting outliers or potential disease indicators without prior labeling.

Despite these efforts, few works have integrated multiple unsupervised algorithms within a single predictive framework for CKD detection. This project addresses this gap by combining K-Means, DBSCAN, Isolation Forest, and Autoencoders, enhanced through feature selection and dimensionality reduction, to build a comprehensive and accurate CKD prediction model.

Table 1. Related Work

Ref	Year	Features	Classifier/Algorithm	Previous Accuracy	Limitations
1	2024	Preprocessed clinical data (missing value imputation, normalization, feature selection)	Various ML models	~86% (varies by model)	graphical interface limits user interaction and makes the system less accessible.
2	2023	Selected medical features for CKD	Supervised learning (e.g., SVM, Decision Trees)	~85% (model-dependent)	Limited labeled data; complexity in medical datasets
3	2023	Engineered features from multiple sources	Ensemble learning (e.g., Random Forest, Gradient Boosting)	88%	Computational complexity; diverse data requirement

III. LITERATURE REVIEW

Ghaniaviyanto Ramadhan et al. (2024) conducted a comprehensive review of machine learning applications in chronic disease prediction, emphasizing the critical role of data preprocessing techniques such as missing value imputation, normalization, and feature selection. Their study highlighted that the quality of data preprocessing directly affects the accuracy and reliability of predictive models in healthcare.

Chaman Deep Kaur et al. (2023) explored various machine learning models specifically for CKD prediction. Their work underlined the significance of selecting the right medical features and classification techniques to improve early diagnosis. The study addressed common challenges such as limited labeled data and the complexity of medical datasets, proposing supervised learning approaches with optimized feature sets. However, the dependence on labeled data presents a limitation in real-world applications, especially in low-resource settings.

Batini Dhanwanth et al. (2023) proposed the use of ensemble learning methods for CKD forecasting. By combining multiple classifiers, their approach enhanced prediction accuracy and reduced overfitting. The paper also highlighted the importance of robust feature engineering and data preprocessing to maximize model performance. While ensemble methods improved reliability, the computational complexity and need for diverse training data remained significant constraints.



IV. METHODOLOGY

We proposed a project to investigate the early prediction and classification of Chronic Kidney Disease (CKD) using unsupervised machine learning techniques. To achieve this, several data preprocessing steps were carried out, including handling missing values, outlier removal, feature normalization, and dimensionality reduction. The proposed framework utilizes algorithms such as K-Means, DBSCAN, Isolation Forest, and Autoencoders to distinguish between CKD and non-CKD cases without relying on labeled training data. A clinical dataset comprising approximately 400 records, including features like blood pressure, serum creatinine, and glucose levels, was used for training and testing the model. This approach aims to support healthcare professionals in identifying CKD at earlier stages, especially in settings where labeled data is scarce or healthcare resources are limited. By leveraging unsupervised learning, the system provides a scalable, non-invasive, and cost-effective solution to assist in timely CKD diagnosis and improve patient outcomes.

V. PROPOSED MODEL

This unsupervised machine learning framework is designed to classify patients into Chronic Kidney Disease (CKD) and non-CKD categories without the use of labeled data. It involves data preprocessing techniques such as handling missing values, normalization, and feature selection to ensure clean and relevant input for the model. The framework utilizes clustering algorithms like K-Means and DBSCAN, along with anomaly prediction methods such as Isolation Forest and Autoencoder, to uncover patterns and outliers in the data. Internal evaluation metrics, including the Silhouette Score, Davies-Bouldin Index, and Reconstruction Error, are used to assess the model's performance. Optionally, validation with labeled data can be conducted to assess accuracy, sensitivity, and specificity. Visualizations like cluster plots, heatmaps, and feature importance charts help in interpreting results, ensuring the model is both lightweight and interpretable for practical use in healthcare settings. Additionally, the framework is scalable, allowing for future development into a web-based diagnostic tool.

VI. IMPLEMENTATION

The dataset employed for this study comprises real-time clinical data sourced from the University of California, encompassing records from 400 patients. The dataset includes 24 key clinical attributes such as blood pressure, blood urea, serum creatinine, glucose level, hemoglobin, and others that are vital for CKD detection. Out of the 400 records, 250 correspond to patients diagnosed with Chronic Kidney Disease (CKD), while 150 are non-CKD cases. This balanced dataset was critical in training and evaluating the models effectively. A comprehensive data preprocessing pipeline was implemented to ensure data quality and improve model performance. This included:

- Data Cleaning: Imputation of missing values using mean, mode, or regression-based techniques.
- Data Transformation: Standardization and normalization (Min-Max Scaling) of continuous features.
- Feature Encoding: Categorical data was converted using one-hot and label encoding methods.
- Feature Selection: Techniques such as Recursive Feature Elimination (RFE), correlation matrix analysis, and SHAP value interpretation were employed to identify and retain the most informative features.

VII. RESULTS

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	36.0	1.2	NaN	NaN	15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	18.0	0.8	NaN	NaN	11.3	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8	NaN	NaN	9.6	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4	NaN	NaN	11.6	35	7300	4.6	no	no	no	good	no	no	ckd

Fig 1.1: Loading data



	age	blood_pressure	specific_gravity	albumin	sugar	red_blood_cells	pus_cell	pus_cell_clumps	bacteria	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	36.0	1.2	NaN	NaN	
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	18.0	0.8	NaN	NaN	
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8	NaN	NaN	
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5	
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4	NaN	NaN	

Fig 1.2: Dropping Id Columns

```
age          0
blood_pressure  0
specific_gravity  0
albumin       0
sugar         0
blood_glucose_random  0
blood_urea     0
serum_creatinine  0
sodium        0
potassium     0
haemoglobin   0
packed_cell_volume  0
white_blood_cell_count  0
red_blood_cell_count  0
dtype: int64
```

Fig 1.3: Random Sampling

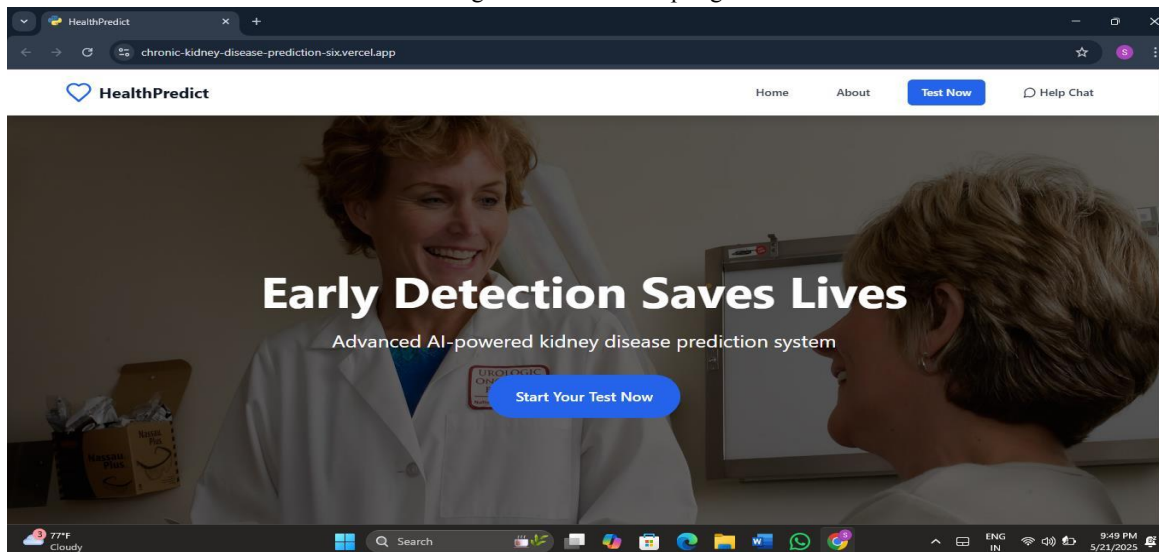


Fig 1.4 : HomePage



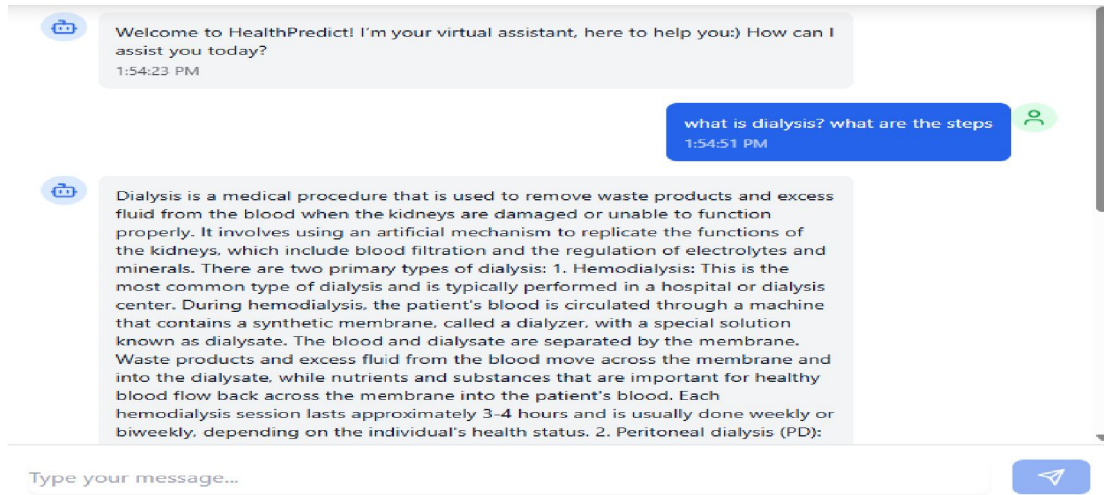


Fig 1.4: AI Assistant

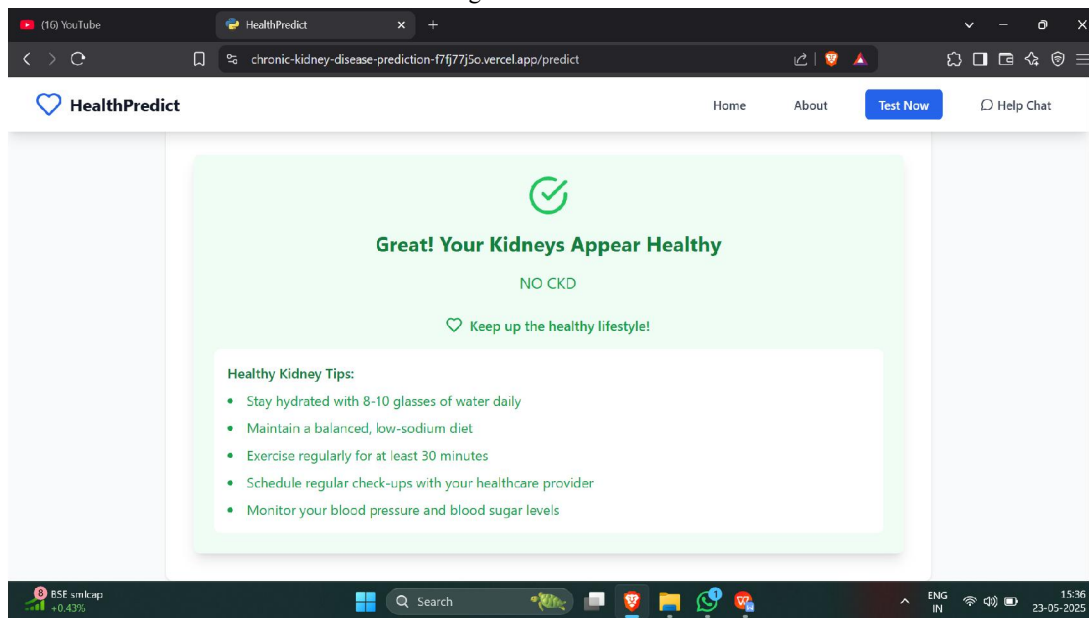


Fig 1.5 : No CKD Result



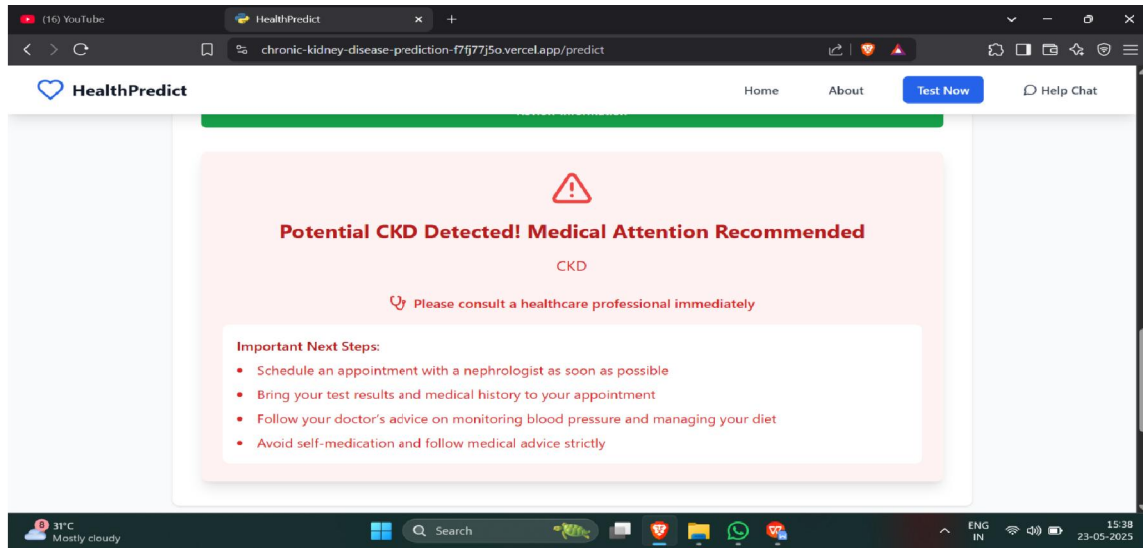


Fig 1.6 : CKD Result

Test Case ID	Scenario	Input	Expected Output
TC01	Valid input data	Patient data (age=45, BP=80, glucose=110)	CKD prediction = "No CKD"
TC02	Missing input data	Patient data with missing glucose level	System replaces missing value with median/mean or raises an appropriate error.
TC03	Invalid input type	Patient data (age="forty-five")	Error message: "Invalid data type for age."
TC04	Extreme input values	Patient data (age=150, BP=300)	System handles gracefully or predicts as "CKD/No CKD" based on thresholds.
TC05	Model accuracy	Known test dataset	Model accuracy >= 90%.
TC06	Scalability	Dataset with 1 million rows	Prediction time <= acceptable threshold (e.g., 2 seconds).
TC07	Data integration	Data from hospital database	System imports data correctly and provides predictions.

Fig 2.1:Test cases are detailed steps to validate the scenarios.



Test Case ID	Test Scenario	Expected Outcome
TC01	Test with missing feature (e.g., age)	Model should either ask for the missing feature or impute a value.
TC02	Test with edge case (e.g., age = 0 or 120)	Ensure the model handles extreme values correctly.
TC03	Test with imbalanced dataset (more healthy than CKD)	Model should still perform well for CKD cases.
TC04	Test accuracy on a separate validation dataset	Accuracy should be above a set threshold (e.g., 90%).
TC05	Test model's response to new, unseen data	The model should adapt and give valid predictions for unseen cases.
TC06	Test with clinically impossible values (e.g., negative blood pressure)	The model should either flag as an error or handle it gracefully.
TC07	Test model on small subgroups (e.g., elderly patients)	Ensure model's predictions are robust for diverse populations.

Fig 2.2 : Test Scenarios for CKD Validation

VIII. CONCLUSION

This work developed an approach for improved prediction and detection of Chronic Kidney Disease based on various unsupervised machine learning approaches including autoencoder, Isolation forest, DB-scan and Kmeans. For considering all the 24 features resulted in a 91% accuracy for I-forest, 94% for DB-Scan, 97.5% for Autoencoder and, 99.3% for K-means clustering. To reduce the time and financial expenses of CHRONIC KIDNEY DISEASE PREDICTION diagnosis, six feature selection strategies, which fall into four distinct categories of feature selection methods, were used. The best features were selected using a set-theory- based rule, which combines multiple feature selection approaches. The data were then classified and validated. For the reduced feature set also Kmeans outperformed other unsupervised algorithms with 99% accuracy. The suggested technique can assist clinicians in managing numerous patients and providing CHRONIC KIDNEY DISEASE PREDICTION diagnoses more quickly. Organizations can use the 53 suggested machine learning architecture in regional clinics with reduced medical expert retention over time, allowing patients in regional locations to receive early diagnosis. As an extension of this work, detection of the five different stages of Chronic Kidney Disease in a similar manner can be done. Thus, would support the medical community in just to detecting the existence of the disease, but also in identifying the stages of the disease.

IX. FUTURE SCOPE

Future enhancements of the proposed model will aim to improve dataset diversity by incorporating plant species from a wider range of geographical regions, including rare and endangered types, and capturing more detailed morphological and visual characteristics to boost identification accuracy in complex environments. Advanced deep learning techniques such as transformers, vision-based attention mechanisms, and hybrid models may be employed to address current limitations like misclassification of visually similar species (e.g., Hibiscus vs. Rose). Environmental factors—such as lighting, soil texture, and climate variability—can introduce noise, requiring refined feature extraction and handling of class imbalance to improve precision and recall. The development of a real-time mobile application for botanists and field researchers is a promising direction, offering on-the-go identification via smartphone cameras with support for



continuous learning. Additionally, integration with GIS platforms for biodiversity mapping and the inclusion of crowdsourced data and AR-based educational tools could significantly enhance scalability, user engagement, and ecological impact.

REFERENCES

- [1]. M. Gjoreski et al., "Chronic Kidney Disease Prediction using a stack of machine- learning classifiers," in 2017 International Conference on Intelligent Environments (IE). IEEE, 2017, pp. 14- 19.
- [2]. J. Voigt et al., "A reevaluation of the costs of heart failure and its implications for allocation of health resources in the United States," Clinical cardiology, vol. 37, no. 5, pp. 312- 321, 2014.
- [3]. G. D. Clifford et al., "Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016," in 2016 Computing in Cardiology Conference (CinC). IEEE, 2016, pp. 609- 612.
- [4]. X. Jiang, Y. Pang, X. Li, and J. Pan, "Speed up deep neural network based pedestrian detection by sharing features across multi-scale models," Neurocomputing, vol. 185, pp. 163- 170, 2016.
- [5]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in International Conference on Neural Information Processing Systems, 2012, pp. 1097- 1105.
- [6]. C. Szegedy et al., "Inception-v4, inception-ResNet and the impact of residual connections on learning," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
- [7]. T. Young et al., "Recent trends in deep learning based natural language processing," IEEE Computational intelligence magazine, vol. 13, no. 3, pp. 55-75, 2018.
- [8]. Y. Bengio et al., "A neural probabilistic language model," Journal of machine learning research, vol. 3, no. Feb, pp. 1137-1155, 2003.
- [9]. S. Amiriparian et al., "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in Interspeech, 2018, pp. 2334-2338.
- [10]. S. Amiriparian et al., "Bag-of-Deep-Features: Noise-robust deep feature representations for audio analysis," in International Joint Conference on Neural Networks (IJCNN). IEEE, 2018, pp. 1- 7.

