



(RESEARCH ARTICLE)



## Chronic kidney disease prediction using machine learning techniques

Ashok Kumar Pasi, Sai Aryan Meesala, Vinesh Doddi, Nandana Ande \*, Thilak Chinta and Nithin Soma

*Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.*

World Journal of Advanced Research and Reviews, 2025, 25(02), 981-989

Publication history: Received on 25 December 2024; revised on 04 February 2025; accepted on 07 February 2025

Article DOI: <https://doi.org/10.30574/wjarr.2025.25.2.0384>

### Abstract

In today's fast-paced world, maintaining health often takes a backseat until visible symptoms arise. Unfortunately, certain diseases, like Chronic Kidney Disease (CKD), develop silently, presenting no noticeable symptoms in the early stages. This delay in detection often leads to severe complications, including kidney failure, cardiovascular disease, or even death. CKD's silent progression highlights the critical need for proactive and predictive healthcare tools that can identify risks early.

Machine Learning (ML) offers a promising solution, capable of analyzing vast amounts of data and predicting potential health risks with high accuracy. In this study, we explored the potential of nine ML techniques for predicting CKD: K-nearest Neighbors (KNN), support vector machines (SVM), logistic regression (LR), Naive Bayes, Extra Tree Classifiers, AdaBoost, XG Boost, and Light GBM. Using a dataset obtained from Kaggle.com with 14 attributes and 400 records related to CKD, we aimed to identify the most effective model for this task.

The attributes included clinical parameters such as blood pressure, specific gravity, albumin, sugar, and more, providing a comprehensive foundation for prediction. Each ML model was meticulously trained and tested, with hyperparameters fine-tuned to achieve optimal performance. Feature scaling and data preprocessing were conducted to ensure the models handled the dataset effectively.

Evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, were used to assess performance.

Among the models, LightGBM emerged as the top performer, achieving an impressive accuracy of 99.00%. This model reformed its counterparts due to its ability to handle imbalanced datasets, fast training speed, and exceptional performance in capturing complex patterns.

**Keywords:** Feature-based sentiment analysis; Customer reviews Support Vector Machines; Term frequency- inverse document frequency

### 1. Introduction

Chronic Kidney Disease (CKD) is a long-term condition characterized by the gradual loss of kidney function over time. The kidneys play a crucial role in filtering waste, excess fluids, and toxins from the blood, maintaining electrolyte balance, and regulating blood pressure. When kidney function declines, harmful substances accumulate in the body, leading to serious health complications.

CKD is commonly caused by underlying conditions such as diabetes, hypertension, and glomerulonephritis. Other risk factors include genetic predisposition, obesity, smoking, and prolonged use of nephrotoxic medications. The disease

\* Corresponding author: Nandana Ande

progresses in stages, ranging from mild kidney impairment to complete kidney failure, requiring dialysis or a kidney transplant.

Symptoms often appear in later stages and include fatigue, swelling in the legs and face, changes in urination patterns, and high blood pressure. Early diagnosis through blood tests, urine tests, and imaging can help slow disease progression with lifestyle modifications, medication, and dietary changes.

CKD is a major public health concern, affecting millions worldwide. Preventive measures such as maintaining a healthy lifestyle, managing chronic conditions, and regular health checkups are crucial in reducing its impact. Increasing awareness and early intervention can significantly improve the quality of life for individuals with CKD.

---

## 2. Related Work

Research on Chronic Kidney Disease (CKD) prediction has evolved significantly with advancements in machine learning, artificial intelligence, and data analytics. Various studies have explored predictive models to detect CKD at an early stage, using patient health records and clinical data.

Several machine learning algorithms, including Decision Trees, Support Vector Machines (SVM), Random Forests, and Neural Networks, have been applied to CKD diagnosis. Studies have shown that ensemble learning techniques and deep learning models improve prediction accuracy by handling complex patterns in medical data. Feature selection methods, such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), have been used to identify the most relevant biomarkers, including creatinine levels, blood pressure, and proteinuria.

Recent research has also integrated real-time health monitoring through wearable devices and Internet of Things (IoT) technologies, enabling continuous tracking of kidney function indicators. Additionally, explainable AI (XAI) techniques have been explored to enhance the interpretability of CKD predictions, helping healthcare professionals understand model decisions.

Despite these advancements, challenges remain in terms of data quality, imbalanced datasets, and generalization across diverse populations. Future research aims to refine predictive models, incorporate multi-modal data, and improve early detection strategies to enhance CKD management and patient outcomes.

Recent advancements include the use of deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, to analyze complex patterns in medical imaging and time-series health data. Furthermore, explainable AI (XAI) is being explored to improve the interpretability of predictions, aiding healthcare professionals in decision-making.

Wearable devices and the Internet of Things (IoT) are also contributing to real-time CKD monitoring, enabling early detection and intervention. Despite these advancements, challenges such as data imbalance, generalization across diverse populations, and privacy concerns persist. Future research aims to refine models, integrate multi-source data, and develop more personalized CKD prediction systems.

Recent studies have explored hybrid models combining traditional statistical methods with deep learning techniques to enhance CKD prediction accuracy. Researchers have leveraged electronic health records (EHRs) and federated learning to develop robust models that ensure data privacy while improving predictive performance. Additionally, natural language processing (NLP) has been applied to extract valuable insights from unstructured clinical notes, aiding in early diagnosis. Transfer learning approaches have also been investigated to adapt pre-trained models for CKD detection with limited datasets. Furthermore, blockchain technology is being explored to secure and decentralize CKD-related health data, ensuring integrity and reducing risks associated with data breaches.

---

## 3. Existing System

The existing system for CKD detection primarily relies on traditional diagnostic methods, including blood tests, urine analysis, and imaging techniques. Physicians manually interpret test results, which can be time-consuming and prone to human error. Some hospitals use basic rule-based systems for early CKD detection, but these lack adaptability to complex patient data.

. Additionally, most CKD diagnoses occur in later stages due to the absence of early warning systems. Limited integration of artificial intelligence (AI) and machine learning (ML) in clinical settings further restricts automated prediction. The existing system lacks real-time monitoring, scalability, and personalized recommendations for early intervention.

Current CKD detection methods often depend on periodic checkups, leading to delayed diagnosis and progression to advanced stages. Many healthcare systems lack centralized patient data, making continuous monitoring difficult. Moreover, limited access to predictive analytics and reliance on subjective clinical assessments reduce early intervention effectiveness, increasing the burden on healthcare infrastructure and patient outcomes.

---

#### 4. Proposed Model

The proposed model leverages machine learning (ML) and deep learning (DL) techniques to enhance early CKD detection and risk assessment. It integrates electronic health records (EHRs), real-time monitoring devices, and patient lifestyle data for comprehensive analysis. The model utilizes feature selection techniques to identify key biomarkers such as creatinine, blood pressure, and protein levels, ensuring high prediction accuracy.

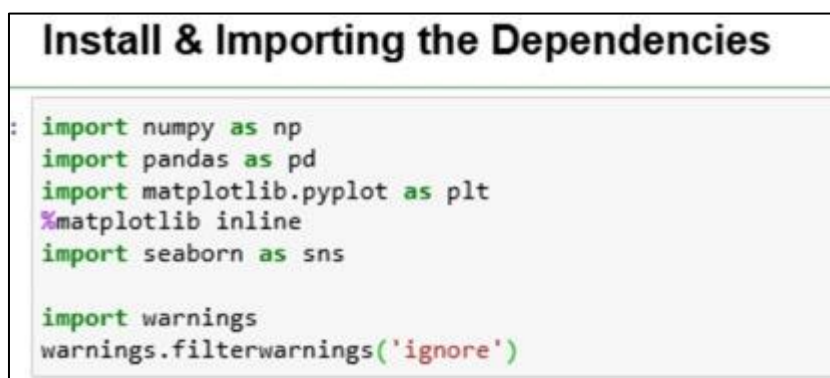
A hybrid AI approach combining Random Forest, Support Vector Machines (SVM), and Neural Networks enhances predictive performance. Additionally, explainable AI (XAI) ensures transparency in decision-making, assisting healthcare professionals in understanding predictions. The system incorporates cloud-based deployment for scalability and enables personalized recommendations based on patient history.

Real-time monitoring through IoT-based wearable devices ensures continuous tracking of kidney function, alerting users to potential risks. Blockchain technology is implemented to secure patient data and ensure privacy. The proposed model aims to provide early diagnosis, improved accuracy, and real-time intervention, reducing CKD progression risks.

---

#### 5. Methodology

The methodology involves data collection from medical records and IoT devices, followed by preprocessing (cleaning, normalization, and feature selection). Machine learning models like Random Forest, SVM, and Neural Networks are trained and validated. Finally, the model is deployed on a cloud-based system for real-time prediction and monitoring.



```

Install & Importing the Dependencies

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')

```

**Figure 1** Libraries imported

Description of the Machine Learning Algorithms and Techniques Chosen: Libraries imported are:

##### 5.1. Data Collection

This is the foundational step where all relevant disease are gathered. The data should be representative of various symptoms.

###### 5.1.1. Dataset Selection

Datasets of kidney disease from Kaggle (e.g., blood pressure and urine test)

###### 5.1.2. Data Preprocessing

The dataset undergoes text cleaning, including removing special characters, stopwords, and redundant whitespace.

Identify relevant biomarkers like creatinine, blood pressure, and protein levels to improve model efficiency. Scale numerical features to ensure uniformity and enhance model performance. Apply techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance CKD and non-CKD cases. Divide the dataset into training, validation, and testing sets to optimize model learning and evaluation.

## 5.2. Pipeline Architecture

Gather patient data from electronic health records (EHRs) and wearable devices, then clean, normalize, and select key features. Apply machine learning algorithms (e.g., Random Forest, SVM, Neural Networks) and optimize using cross-validation. Use trained models to predict CKD risk and apply explainable AI (XAI) for transparent decision-making. Deploy the model on a cloud-based system with IoT integration for real-time monitoring and continuous updates.

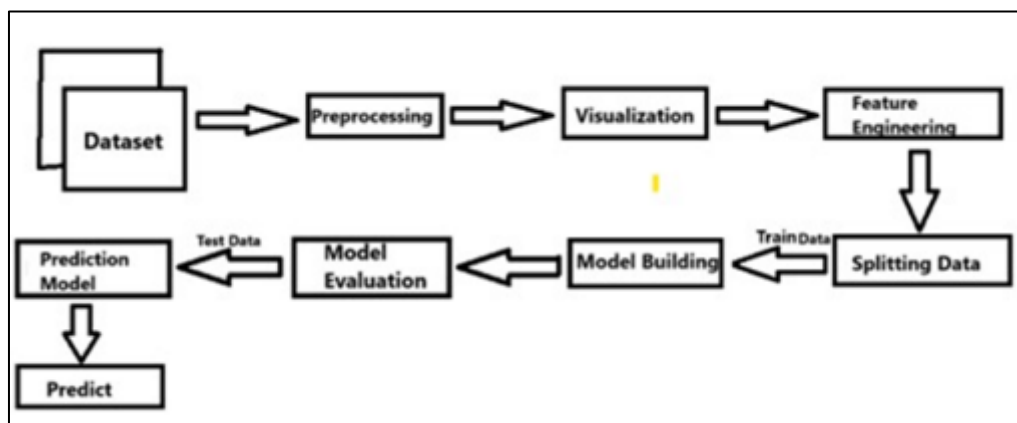


Figure 2 Pipeline Architecture

### System Architecture Components

- **Input Capture:** The system captures user input features such as age, sex, blood pressure, blood sugar levels, creatinine levels, proteinuria, and lifestyle factors through the Streamlit interface.
- **Data Preprocessing:** The system processes the captured inputs by normalizing numerical values, handling missing data, and encoding categorical variables to prepare them for prediction.
- **CKD Prediction:** A trained Machine Learning model (e.g., Logistic Regression, Random Forest) predicts the likelihood of Chronic Kidney Disease based on the input features.
- **Confidence Scoring:** The system calculates a confidence score for the prediction, reflecting the reliability of the model's output.
- **Display Results:** The results (predicted CKD risk and confidence scores) are displayed on the user interface (UI) in real-time.
- **Real-Time Update:** The process dynamically updates as new inputs are provided, with updated risk predictions and scores shown on the UI.
- **Output:** The final output is a clear and intuitive display of the predicted CKD risk, accompanied by recommendations for further action (e.g., consulting a healthcare provider).

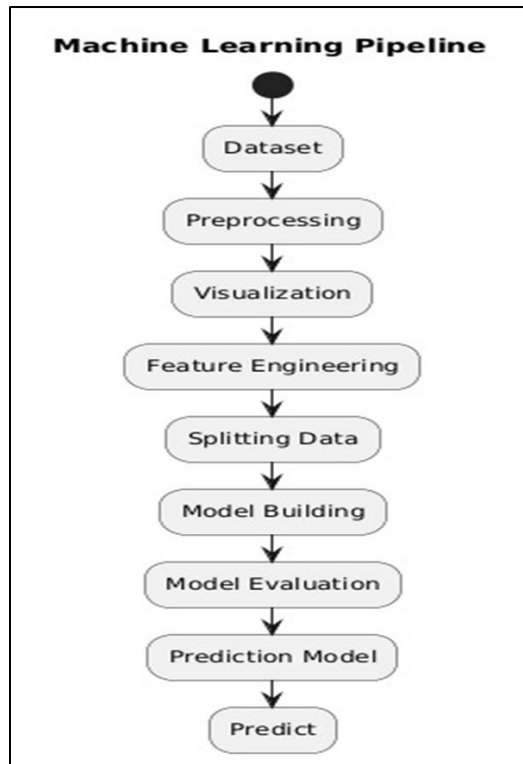
This workflow ensures the system processes inputs, predicts insurance costs, and provides real-time feedback efficiently

## 5.3. Model Development

The model is the core component of the system, responsible for learning and prediction.

### 5.3.1. ML Model Architecture:

For a Chronic Kidney Disease (CKD) detection and prediction system using Machine Learning (ML), the model architecture can be designed based on structured clinical data (e.g., lab test results, patient history) and unstructured data (e.g., medical notes). Below is a detailed ML model architecture for CKD prediction



**Figure 3** ML Model Architecture

**5.3.2. Training:**

Train the model using Train test split function

```

109]: #calling the XGBClassifier model
classifier=XGBClassifier()

# Training the model
classifier.fit(X_train,y_train)

[16:17:37] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.4.0/src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

109]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
  colsample_bynode=1, colsample_bytree=0.3, gamma=0.2, gpu_id=-1,
  importance_type='gain', interaction_constraints='',
  learning_rate=0.300000012, max_delta_step=0, max_depth=5,
  min_child_weight=1, missing=nan, monotone_constraints='()',
  n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=0,
  reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
  tree_method='exact', validate_parameters=1, verbosity=None)
  
```

**Figure 4** Train test split function

**5.3.3. Testing**

Evaluate the model on unseen test data to ensure its ability to generalize to new inputs. Use performance metrics such as accuracy, precision, recall to assess effectiveness.

```

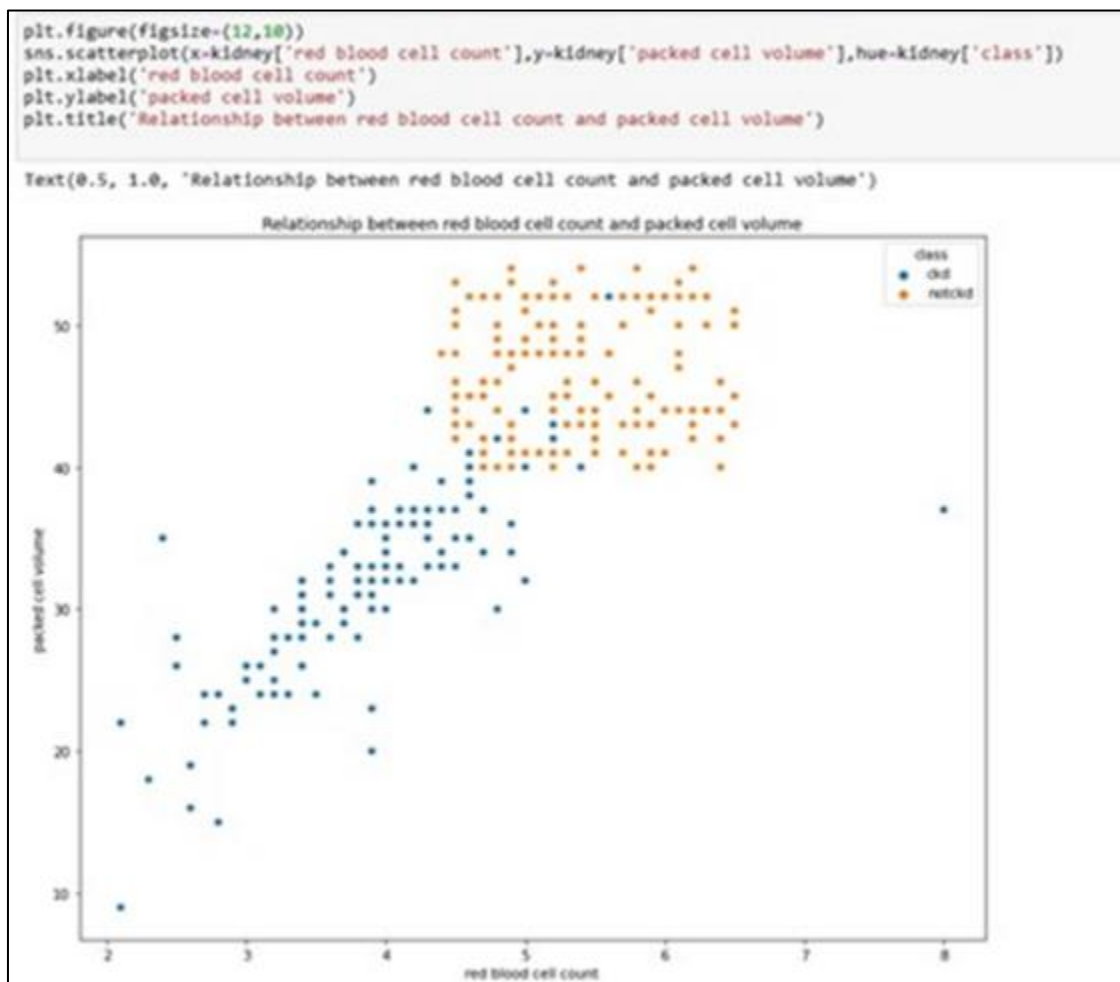
In [96]: from sklearn.model_selection import train_test_split
In [97]: X_train,X_test,y_train,y_test=train_test_split(X_new,y,random_state=0,test_size=0.3)
In [98]: X_train.shape
Out[98]: (280, 10)
In [99]: y_train.value_counts() #Checking for imbalancing
Out[99]: 0    178
         1    102
         Name: class, dtype: int64

```

**Figure 5** Model Training/Testing

#### 5.4. Visualization

This involves presenting visual reports, such as sentiment distribution graphs, word clouds, and feature-specific insights, to effectively communicate the results. Additionally, actionable insights are provided, highlighting key areas for improvement based on customer feedback



**Figure 6** Visualization

## 6. Results and Discussion

age	bp	al	su	rbc	pc	pcr	fe	bgr	bu	ur	gpt	wt	hba	dm	cad	pe	ure	Disease
24	120	2	0	1	0	1	0	130	80	1.9	3.7	9000	1	1	0	0	1	Present
48	80	3	0	0	1	0	0	137	162	0.6	4.9	11000	0	1	0	0	1	Present
51	0	0	0	1	0	0	0	121	27	0.8	3.7	8000	0	0	0	0	0	Healthy

Figure 7 kidney disease prediction

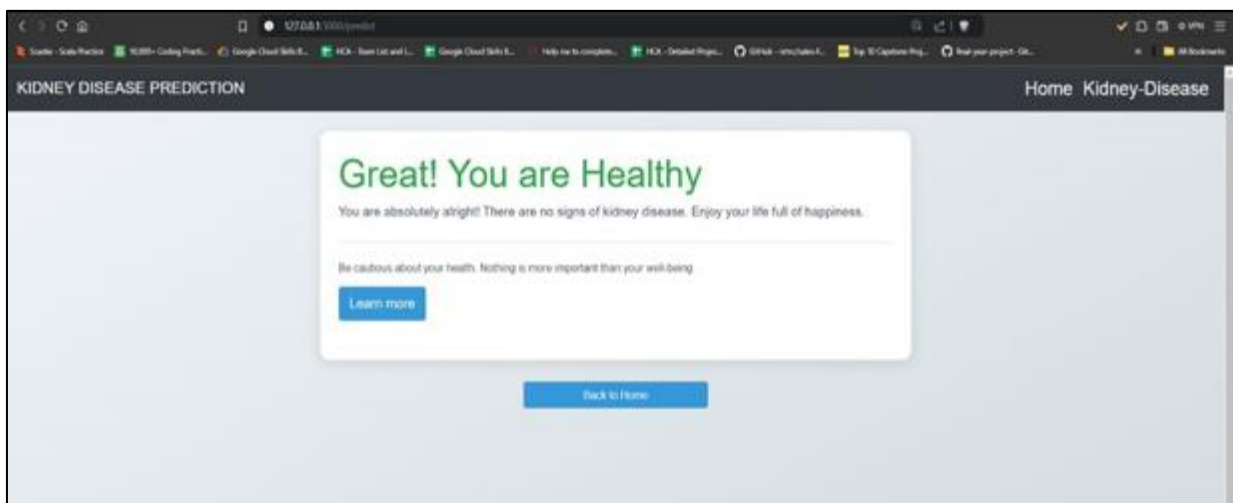


Figure 8 Positive Prediction

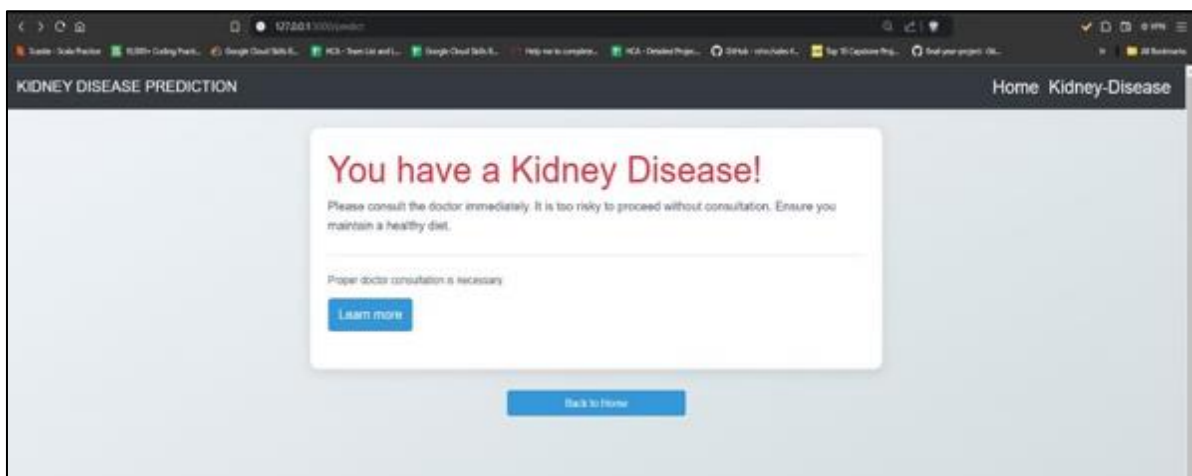


Figure 9 Negative Prediction

## 7. Conclusion

Building a Chronic Kidney Disease (CKD) detection and prediction model requires a systematic approach that ensures high accuracy, robustness, and real-world applicability. The key takeaways from model design and validation process are: Feature Selection & Data Processing: Proper feature, Model Selection, Validation for Reliability, Handling Imbalanced Data, Deployment & Monitoring. A well-validated CKD prediction model can assist doctors in early detection, risk assessment, and treatment planning, ultimately improving patient outcomes. By continuously refining the model and integrating real-time monitoring, the system can evolve into a powerful AI-driven healthcare assistant.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References

- [1] KUMAR, P. ASHOK, G.Satish KUMAR, and SETTI NARESH KUMAR. "Improve the Capacity of Uniform Embedding for Efficient JPEG Steganography Based on DCT." (2015).
- [2] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In XVIII International Conference on Data Science and Intelligent Analysis of Information, pp. 13-26. Cham: Springer Nature Switzerland, 2023.
- [3] Sunkavalli, Jayaprakash, B. Madhav Rao, M. Trinath Basu, Harish Dutt Sharma, P. Ashok Kumar, and Ketan Anand. "Experimentation Analysis of VQC and QSVM on Sentence Classification in Quantum Paradigm." In 2024 International Conference on Computing, Sciences and Communications (ICCS), pp. 1-5. IEEE, 2024.
- [4] Kumar, P. Ashok. "Event Based Time Series Sentiment Trend Analysis."w
- [5] Kumar, P. Ashok. "Event Based Time Series Sentiment Trend Analysis."
- [6] PANDI CHIRANJEEVI, THATIKONDA SUPRAJA, P. ASHOK KUMAR, and RALLA SURESH. "A SURVEY: RECOMMENDER SYSTEM FOR TRUSTWORTHY."
- [7] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Correction to: Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In XVIII International Conference on Data Science and Intelligent Analysis of Information, pp. C1-C1. Cham: Springer Nature Switzerland, 2023.
- [8] Centers for Disease Control and Prevention, "National Chronic Kidney Disease Fact Sheet, 2017," 2017.
- [9] Centers for Disease control and Prevention, "chronic kidney disease in the United States, 2021," 2021.
- [10] Levey A. S. and Coresh J., "chronic kidney disease," Lancet, vol. 379, no. 9811, pp. 165–180, 2012. doi: 10.1016/S0140-6736(11)60178-5 [DOI] [PubMed] [Google Scholar]
- [11] Kovesdy C. P., "Epidemiology of chronic kidney disease: an update 2022," Kidney International Supplements, vol. 12, no. 1, pp. 7–11, 2022. doi: 10.1016/j.kisu.2021.11.003 [DOI] [PMC free article] [PubMed] [Google Scholar]
- [12] KDIGO, "kidney disease: Improving Global Outcomes (KDIGO) CKD Work Group KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease," Kidney International Supplements, vol. 3, no. 1, 2013.

## Author's short biography

Mr P Ashok Kumar

I'm Mr. P. Ashok Kumar, working as Assistant Professor in Computer Science and Engineering (Data Science) at ACE Engineering College, Hyderabad, Telangana, Having 15+ years of teaching experience and one year in the industry. Holding a B.Tech, M.Tech, and Ph.D., my research focuses on Machine Learning and Deep Learning. I aim to inspire students and contribute to advancements in technology through my work.





<p><b>M Sai Aryan:</b>          I am a B.Tech student with a strong interest in Blockchain and Machine Learning. Currently, I am expanding my skills in Full Stack Development and Blockchain technologies. My research focuses on Blockchain applications and Machine Learning, aiming to build innovative solutions that leverage these technologies. I am passionate about exploring decentralized systems and their integration with AI and data analytics to solve complex problems. As part of the Feature-Specific Sentiment Analysis project, I applied machine learning techniques to extract insights from customer feedback. This contributed to feature-based sentiment categorization, enhancing the decision-making process and providing valuable information for product development.</p>	
<p><b>D Vinesh:</b>          I am pursuing my B.Tech in Data Science, with experience in Machine Learning, Artificial Intelligence, and Data Analytics. My research interests include sentiment analysis, computer vision, and AI-powered decision systems. I have contributed to projects like Feature-Based Sentiment Analysis, virtual try-on systems, and early pest detection using AI. By applying machine learning and deep learning techniques, I aim to develop impactful solutions to real-world problems. My focus is on leveraging AI technologies to enhance decision-making and improve operational efficiency across various industries, combining theory with practical applications for positive outcomes.</p>	
<p><b>A Nandana:</b>          I am a B.Tech student with a passion for software development, automation, and data analytics. My expertise includes Java development, Python programming, Power BI, and UiPath for process automation. I have hands-on experience automating workflows with UiPath, improving operational efficiency. One of my significant projects is "Specific Sentiment Analysis of iPhone Reviews," where I used machine learning to extract insights from customer feedback. I also utilize Power BI for data visualization, aiming to turn complex datasets into actionable insights and help businesses make informed decisions.</p>	
<p><b>Ch Thilak:</b>          I am a B.Tech student with a keen interest in Convolutional Neural Networks (CNN) and Artificial Intelligence. My focus is on advancing AI knowledge, especially in computer vision applications. I am passionate about solving real-world problems using CNN techniques in image recognition, object detection, and other AI-driven tasks. I contributed to the Feature-Specific Sentiment Analysis project, applying machine learning models to analyze and categorize customer feedback based on product features, providing valuable insights into user sentiments and enhancing the product development process.</p>	
<p><b>S Nithin :</b>          I am a B.Tech student with a passion for software development, automation, and data analytics.          My expertise includes Java development, Python programming, Power BI, and UiPath for process automation. My research interests lie in sentiment analysis, process automation, and data-driven decision systems. I have hands-on experience automating workflows with UiPath, improving operational efficiency. One of my significant projects is "Specific Sentiment Analysis of iPhone</p>	