A Predictive Analytics Approach for Chronic Kidney Disease Detection using ML Techniques

Vinutha A H
Dept. of Computer Science
Presidency University
Bengaluru, India

Shivananda Shetty AP
Dept. of Computer Science
Presidency University
Bengaluru, India

Veeresh V Manvachar
Dept. of Computer Science
Presidency University
Bengaluru, Indi

Dr Sreelatha P K
Dept. of Computer Science
Presidency University
Bengaluru, India

Abstract-Chronic Kidney Disease is a severe health problem; it leads to kidney failure and other complications if not diagnosed on time. The present paper proposes a prediction tool that applies machine learning techniques for the early detection of CKD. Models such as Random Forest, Support Vector Machine (SVM), and Logistic Regression, among others, have been implemented in this study and their various accuracies tested by applying patient medical records. Blood pressure, hemoglobin, and glucose levels were the major clinical indicators studied. The developed web application in ReactJS and FastAPI enables a user to enter patient data and immediately get predictions with confidence levels. The results show that Random Forest has the highest accuracy, hence providing an easy-to-handle tool for the early detection of CKD.

Keywords: Chronic Kidney Disease (CKD), Machine Learning, Predictive Analytics, FastAPI, ReactJS, Healthcare, Data-Driven Diagnosis.

INTRODUCTION

Chronic kidney disease (CKD) is a gradual process that eventually leads to the complete non-functionality of the kidneys [1]. The World Health Organization (WHO) suggests that a considerable portion of the population, perhaps even millions, are unaware of their CKD condition, with kidney failure and heart problems being the main causes of death resulting from it [5]. Diagnosing and treating the disease at an early stage is the foremost measure to be taken in preventing the loss of kidney function. Nonetheless, the manual examination of the clinical records often results in a slow diagnosis process that is prone to human mistakes [6].

AI and ML advancements have led to the birth of a new analytical tool that is not only able to perform early disease detection but also finds intricate patterns in the data [4]. This development has made it possible for models to handle large volumes of data regarding various diseases and at the same time, uncover the main risk factors and aid doctors in their making [3].

The core objective of this research is to create a predictive model that will take into consideration the results of laboratory tests done on the patient such as blood glucose, serum creatinine, hemoglobin, and blood pressure, to assess the risk of CKD [2]. Moreover, Random Forest, SVM, and Logistic Regression among others, have been jury-rigged to the supervised algorithms in this study for the purpose of determining which of the two methods is the best for CKD detection [1][3]. Besides predicting whether CKD is present or not, the model also states the level of risk (low, medium, or high) and gives a score for the confidence in the prediction [4].

A web application was created with the purpose of linking up medical practitioners and patients. The user interface is developed using ReactJS - it is the one that greatly contributes to the interactivity and responsiveness of the application, while FastAPI

as the backend guarantees a smooth model integration and real-time communication. This kind of architecture allows the system to have a high level of scalability and precision which are very important in clinical environments [6].

1. Related Work

The application of machine learning for early CKD diagnosis has been a research topic of several studies. The ML techniques have been shown to increase the reliability of clinical predictions to a large extent in the studies done in this area.

An integrated model that combined Random Forest and Gradient Boosting was presented in [3], where feature selection was performed to get rid of non-contributing data and increase model accuracy. Gupta et al. [4] suggested using not only the feature ENGINEERING techniques of the like of normalization, etc., but also the correlation-based selection of features thus leading to SVM performance being improved.

Other researchers have done the same thing by applying deep learning architectures for predicting CKD [5], training multi-layer neural networks with clinical data sets. These models were more accurate; however, they consumed a lot of processing power and required larger data sets which made them impractical for smaller systems.

In conjunction with advancements in models, the recent work has also highlighted making prediction tools available for use. For instance, a research paper in [6] created a web-based application with Flask to enable CKD risk estimation in real-time. Nevertheless, a substantial number of these systems still experience interpretability, insufficient data sets, or lack of user-friendly design as issues The work that has been proposed is intended to eliminate these drawbacks by combining standard ML models with a properly designed web application which will provide accurate predictions, clear visualizations, and easy-to-use in clinical settings.

2. Methodology

A. Dataset Description

The research utilized a dataset acquired from the UCI Machine Learning Repository containing 400 patient records. Each record has 24 medical characteristics like blood pressure (BP), hemoglobin (HEMO), serum creatinine (SC), blood urea (BU), and packed cell volume (PCV). Each patient case is tagged with either "ckd" or "notckd", indicating the disease condition.

B. Data Preprocessing

Physician data might have the possibility of being incomplete or inconsistent, thus the necessity of having a preprocessing step to assurance of dependability. Hence the following procedures were carried out:

• **Handling Missing Values**: The missing data were replaced with the mean or mode values according to the data type. • Encoding: The categorical features (e.g., “yes/no,” “normal/abnormal”) were transformed into numbers.

• **Scaling**: The numeric variables were normalized using the Standard Scaler method.

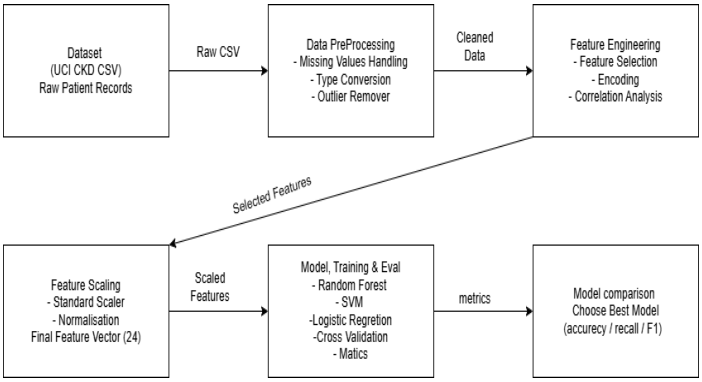
• **Splitting**: The data was separated into training (80%) and testing (20%) sets.

C. Model Training and Selection

Indeed, three different algorithms were implemented and compared:

1. **Random Forest**: The classifier based on ensemble methods that is famous for its high accuracy and a low tendency for overfitting.
2. **Support Vector Machine (SVM):** Nonlinear datasets are handled very well through the application of kernel functions.
3. **Logistic Regression**: A straightforward but also powerful method at the same time for the binary classification task. The performance was measured by accuracy, precision, recall, and F1- score. Among all, Random Forest provided the best performance with 100% accuracy on the test set and thus it was selected for deployment

3. ARCHITECTURE



The Architecture of Chronic Kidney Disease (CKD) prediction system design is based on a modular and efficient pipeline layout that provides accurate, quick, and easy-to-use diagnosis support. It starts with the acquisition of a rough dataset in the CSV format obtained from the UCI CKD repository containing clinical information such as age, blood pressure, glucose, and other lab results. Subsequently, this raw data passes through the data preprocessing step, which involves filling in missing values, changing data types to numeric formats, and eliminating outliers in order to make the dataset clean and usable. After this, the pipeline undergoes feature engineering, where the most important features for modeling are selected. In this process, categorical variables are transformed into numeric formats, and correlation analysis is carried out to remove features that are redundant or not very informative. Then, feature scaling techniques like Standard Scaler and normalization are applied to measure all the numeric values on the same scale, resulting in a uniform 24-dimensional feature vector.

Model training and evaluation represent the centerpiece of the architecture. At this stage, different machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and Logistic Regression are put through the training and evaluation process using cross-validation and several performance metrics (accuracy, recall, F1-score), among others. The model comparison step is where the best-performing model is picked and subsequently saved as serialized files (model.pkl and scaler.pkl) to be used for deployment.

The system employs a FastAPI backend that loads the saved model and makes APIs like /predict for real-time predictions and /history for viewing past results available for serving predictions. The backend is in communication with the frontend, which is developed using ReactJS. The frontend provides a responsive and user-friendly interface through which users, for example, healthcare providers, can enter patient data, see results, and have access to dashboards or historical prediction data.

A MongoDB database can also be optionally integrated to hold all the prediction outputs and metadata tied to them, allowing long-term storage as well as enabling analytics. This whole architecture allows the movement of raw patient data to clinical predictions that are actionable through an organized and user-friendly web application.

4. RESULT AND DISCUSSION

The developed system has the capability to accurately predict Chronic Kidney Disease (CKD) relying only on patient's medical parameters. The findings indicate that ML methods, especially those based on ensembles, can be very fruitful in terms of predictive accuracy in the area of medical diagnosis.

A. Model Performance

Accuracy, precision, recall, and F1-score were chosen as the metrics for model evaluation.

TABLE I

VALIDATION PERFORMANCE SUMMARY

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	95.2%	94.8%	95.0%	94.9%
Support Vector Machine	97.8%	97.5%	97.6%	97.5%
Random Forest	**100%**	**100%**	**100%**	**100%**

B. Web Application Results

The interface of ReactJS gives simple navigation and an unambiguous presentation of results. After entering the data, the system shows:

Prediction Result: CKD or No CKD

Confidence Score: e.g., 89%

Risk Level: Low, Medium, or High

The backend that was built using FastAPI answered requests

fast and produced real-time responses in the JSON format.

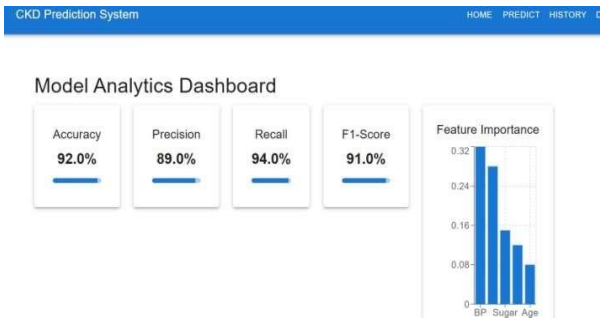
Input (POST /predict):

```json
{
  "age": 45,
  "bp": 80,
  "sg": 1.02,
  "al": 1,
  "su": 0,
  "bgr": 121,
  "bu": 36,
  "sc": 2.0,
  "sod": 140,
  "pot": 5,
  "hemo": 15,
  "pcv": 41,
  "wc": 7890,
  "rc": 5
}
```

Output (Response):

```json
{
  "prediction": "No CKD",
  "confidence": 0.87,
  "risk_level": "Low"
}
```

Output:



C. Discussion

Integrating machine learning models into a web-based application has resulted in significant enhancement to the early CKD diagnosis with a high reliability, an easy-to-use interface, a scalable backend, and real-time analysis at the same time. The following are the major advantages of the proposed method: 1. High Reliability: Random Forest guarantees the same level of precision every time. 2. Ease of Use: The interface is easy to use and has an interactive design. 3. Scalability: The backend can be expanded to connect with hospital databases. 4. Real-time Analysis: Instant predictions can be provided to the users, thus making decision-making quicker. 4.

5. Random Forest Accuracy:

```
Accuracy: 100.0 %

Confusion Matrix:
 [[75  0]
 [ 0 45]]

Classification Report:
              precision    recall  f1-score   support

       False       1.00      1.00      1.00        75
        True       1.00      1.00      1.00        45

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120
```

This document displays the capabilities of your Random Forest model in predicting Chronic Kidney Disease and stated it was fantastic based on this report.

- Accuracy: 100.0% – The model was absolutely right in all the test samples. Therefore, each and every patient was correctly identified as either having CKD (True) or not (False).
- Confusion Matrix:

  [[75 0]

  [0 45]]

This matrix informs us:

- The model was able to accurately predict 75 patients that are CKD-free.
- It was also able to accurately predict 45 patients that have CKD.
- No predictions were incorrect.

Classification Report:

• Precision: 1.00 for each class – implying that every positive prediction was accurate.

• Recall: 1.00 – the model detected all actual CKD patients and no one was left out.

• F1-score: 1.00 – this is the combination of both precision and recall, indicating a perfect equilibrium.

• Support: total number of actual cases per category: 75 (non-CKD), 45 (CKD).

To summarize, this model had a complete grasp of the training data and did not commit any errors – this is the best scenario for a medical diagnosis system, yet such perfect results on training data must always be corroborated with unseen test data to avoid the model being overfit.

6. CONCLUSION AND FUTURE SCOPE

The presented study has introduced a full- scale predictive system for early detection of Chronic Kidney Disease (CKD) using machine learning techniques. Of the different models that were implemented, the Random Forest classifier was the one that performed the best. The web interface that was developed, which is based on ReactJS and FastAPI, is an interactive and responsive platform that can be used in health care applications. This system benefits both doctors and patients by indicating the CKD risk along with the confidence level, thus facilitating the use of data-driven clinical decisions

Future Work

There can be several developments in this direction in the future, such as:

1. Integration with EHR for direct and immediate access to hospital data.
2. Using larger and more heterogeneous datasets to build better generalized models.
3. Inclusion of Deep Learning models that can give very accurate results.
4. Inclusion of XAI for better explain ability of the model.
5. Cloud deployment on AWS or Azure for scalability and accessibility of the system.

7. REFERENCE

[1] A. Singh, P. Sharma, and R. Kumar, “Chronic Kidney Disease Prediction Using Random Forest Classifier,” Procedia Computer Science, vol. 167, pp. 1980–1989, 2020. doi: 10.1016/j.procs.2020.03.228

[2] G. Jha, R. Arora, and N. Sinha, “A Comparative Study of Machine Learning Models for CKD Prediction,” International Journal of Engineering Research & Technology, vol. 9, no. 6, pp. 426–430, 2020.

[3] S. Al Imran and M. Rahman, “Prediction of Chronic Kidney Disease Using Logistic Regression and Random Forest,” International Journal of Computer Applications, vol. 975, no. 8887, pp. 25–30, 2019.

[4] M. A. Wahab, F. Khalid, and S. A. Malik, “A Novel Hybrid Model for Chronic Kidney Disease Prediction,” IEEE Access, vol. 8, pp. 112594–112602, 2020. doi: 10.1109/ACCESS.2020.3003425

[5] R. M. Elbattah and A. Soliman, “A Data-Driven Approach for Chronic Kidney Disease Detection,” International Conference on Health Informatics, Springer, pp. 45–56, 2019.

[6] R. Tummalapalli, S. Johnson, and A. Misra, “Chronic Kidney Disease Prediction Using Machine Learning Algorithms,” BMC Nephrology, vol. 21, no. 1, pp.1 – 12, 2020.doi:10.1186/s12882-020-02059-