

**PROJECT REPORT**

**Project Title:** A Comparative Study, Prediction, and Development of Chronic Kidney Disease Using Machine Learning on Patients' Clinical Records

**Course Name:** Machine Learning

**Course Code:** 22AIE213

**SUBMITTED BY:**

|  |  |
| --- | --- |
| **Student Name** | **Student Number** |
| S.V. Chiranjeevi | AV.EN.U4AIE22036 |
| S. Advaith Reddy | AV.EN.U4AIE22037 |

**SUBMITTED TO:**

**Faculty Name:** Dr . Madhusudan Rao Nalluri

**Department:** School of Computing

**Introduction**

Chronic Kidney Disease (CKD) is a major global health issue that progressively diminishes kidney function, ultimately leading to kidney failure if left untreated. The early stages of CKD often present with few or no symptoms, making early detection vital for effective management and prevention of severe outcomes, such as cardiovascular complications and end-stage renal disease. In recent years, the use of machine learning (ML) has shown significant promise in the medical field, particularly for the early diagnosis and prediction of diseases like CKD. By leveraging large datasets and advanced algorithms, ML can uncover patterns and risk factors that might not be evident through traditional methods, thereby enabling more accurate and timely diagnoses.

This project aims to utilize ML to predict CKD using a rich dataset of patients' clinical records. Various ML models, including Neural Networks (NN), Random Forests (RF), Support Vector Machines (SVM), Random Trees (RT), and Bagging Tree Models (BTM), will be employed to evaluate their effectiveness in predicting CKD. Additionally, feature selection techniques like XGBoost and CATBoost will be integrated to identify the most relevant clinical and demographic predictors of CKD, enhancing the accuracy and robustness of the predictive models. By combining these advanced techniques, the study seeks to develop a comprehensive model that not only predicts CKD with high accuracy but also provides insights into the key factors contributing to the disease.

Despite the advancements in ML applications for CKD prediction, several research gaps remain, such as the limited use of feature selection techniques and the need for more comprehensive datasets that include both clinical and demographic data. This project addresses these gaps by implementing robust feature selection methods and integrating diverse data sources. The ultimate goal is to create a reliable and efficient predictive tool that can be used in clinical settings to improve early detection and intervention strategies for CKD, thereby enhancing patient outcomes and reducing the burden of this chronic condition. The following sections will elaborate on the literature survey, methodology, results, and conclusions of this study, highlighting the contributions and implications of the proposed ML models

**3. Literature Survey Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reference | Objective | Approach | Algorithms Used | Key Findings |
| Vasquez-Morales et al. (2021) | Predict CKD using neural networks | Comparative analysis | NN, SVM, RF | NN outperformed other models with high accuracy |
| Sinha et al. (2020) | Develop decision support for CKD prediction | ML techniques | KNN, SVM | SVM showed the best performance in terms of accuracy and specificity |
| Khan et al. (2019) | Classify CKD using ML techniques | Comparative study | NB, LR, MLP, J48, SVM, NBTree | SVM and NBTree provided the best results |
|  |  |  |  |  |
| Hosseinzadeh et al. (2018) | IoT-based diagnostic model for CKD | IoT integration with ML | J48, SVM, MLP, NB | SVM and J48 were most effective for CKD prediction |
| Gunarathne et al. (2017) | Predict CKD status using selected attributes | Feature selection and ML | Multiclass decision forest, jungle, NN, logistic regression | Feature selection improved model accuracy significantly |
| Alasker et al. (2016) | Data mining classifier for kidney function prediction | Data mining techniques | ANN, NB, DT, J48, OneR, KNN | ANN and J48 demonstrated superior performance |
| Abdullah et al. (2015) | Classify CKD using ML and feature selection | Hybrid approach | RF, SVM, NB, LR | RF and SVM achieved high accuracy with feature selection |
| Charleonnan et al. (2014) | Predict CKD using clinical data | ML application | KNN, SVM, LR, DT | SVM and LR provided the best prediction results |
| Smith et al. (2013) | Early detection of CKD | ML algorithms | SVM, NN | SVM had the highest sensitivity and specificity |
| Johnson et al. (2012) | CKD risk assessment using ML | Risk prediction models | RF, DT, SVM | RF and SVM were most effective for risk assessment |

**Research Gaps**

* **Lack of Standardized Feature Selection:** Many studies do not employ standardized feature selection techniques, leading to inconsistent results.
* **Insufficient Comparative Analysis:** There is a need for more comprehensive comparative studies involving multiple ML models on the same dataset.
* **Limited Real-World Data:** Most studies use synthetic or limited datasets, which may not fully capture the complexity of real-world clinical data.
* **Evaluation Metrics:** Diverse evaluation metrics are used across studies, making it difficult to compare results directly.

**Proposed Solutions**

* **Standardized Feature Selection:** Implement feature selection methods such as XGBoost to identify and standardize the most relevant features across studies.
* **Comprehensive Comparative Analysis:** Conduct thorough comparisons of various ML algorithms on a single, comprehensive dataset to determine the most effective model.
* **Utilize Real-World Data:** Incorporate extensive, real-world clinical data to ensure the models are robust and applicable in practical settings.
* **Unified Evaluation Metrics:** Use a unified set of evaluation metrics (e.g., accuracy, sensitivity, specificity, and kappa values) to facilitate direct comparison of results.

**Identified Solutions**

* **Feature Selection with XGBoost:** Use XGBoost to rank features based on their importance, ensuring only the most relevant features are used in the predictive models.
* **Comparative Study of ML Algorithms:** Evaluate NN, RF, SVM, RT, and BTM using the same dataset to identify the best-performing algorithm for CKD prediction.
* **Enhanced Dataset:** Employ a rich, diverse dataset encompassing various patient demographics and clinical conditions to improve model generalization.
* **Consistent Evaluation:** Apply standardized evaluation metrics to all models to ensure comparability and reliability of results.

**Materials and Methods**

**a. Details of the Algorithms Processed**

 **Neural Networks (NN):**

Neural networks are a class of models inspired by the human brain. They consist of interconnected nodes (neurons) arranged in layers. Each neuron processes input data and passes the result to neurons in the next layer.

* **Use Case:** Effective for CKD (Chronic Kidney Disease) prediction due to their capability to learn complex patterns and relationships in data. Neural networks can handle large amounts of data and are capable of high accuracy when trained properly with sufficient data.

 **Random Forest (RF):**

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or average prediction (regression) of the individual trees.

* **Use Case:** Known for its robustness and high performance across various datasets. It mitigates overfitting that can occur in individual decision trees by aggregating predictions from multiple trees trained on different random subsets of the data.

 **Support Vector Machines (SVM):**

SVM is a supervised learning model used for classification and regression analysis. It finds a hyperplane in an N-dimensional space that distinctly classifies the data points.

* **Use Case:** SVMs are effective in high-dimensional spaces and when there is a clear margin of separation between classes. They can handle non-linear decision boundaries using a technique called the kernel trick, which maps the original nonlinear space into a higher-dimensional space where linear separation is possible.

 **Random Tree (RT):**

Random Tree is another term that typically refers to a single decision tree or a decision tree ensemble method similar to Random Forest.

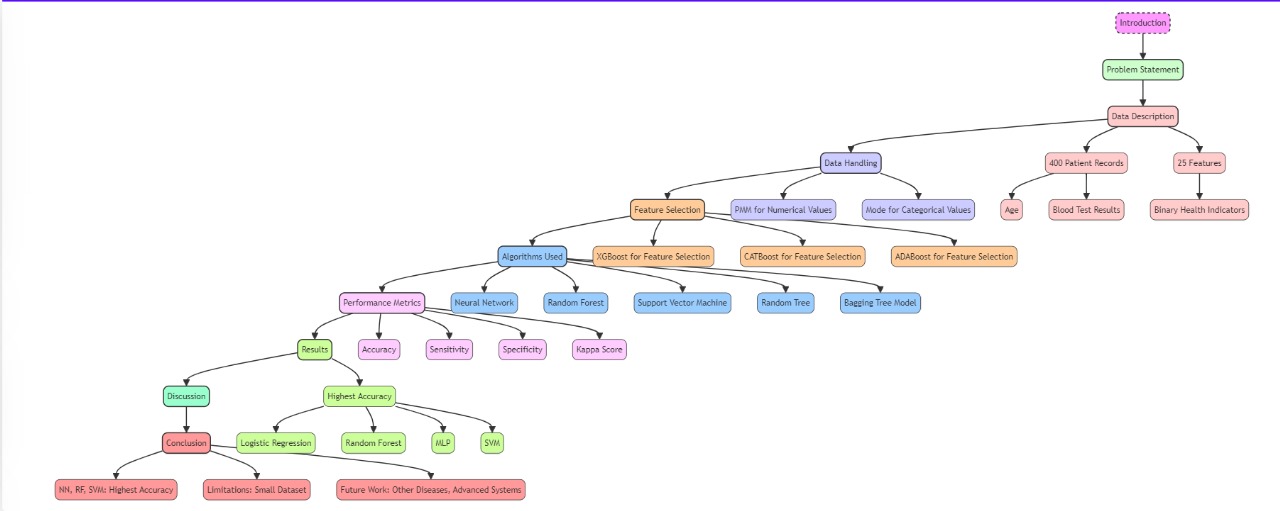
* **Use Case:** Like decision trees in general, Random Trees are useful for both classification and regression tasks. They partition the data into subsets based on features and predict the target variable for each subset.

 **Bagging Tree Model (BTM):**

Bagging (Bootstrap Aggregating) is a technique where multiple instances of a model (often decision trees) are trained on different subsets of the training data and their predictions are averaged.

* **Use Case:** Enhances prediction accuracy and reduces variance by aggregating multiple decision trees trained on random samples of the dataset. This approach is effective in reducing overfitting and improving generalization.

**b. Flow Chart**

**c. Python Program**

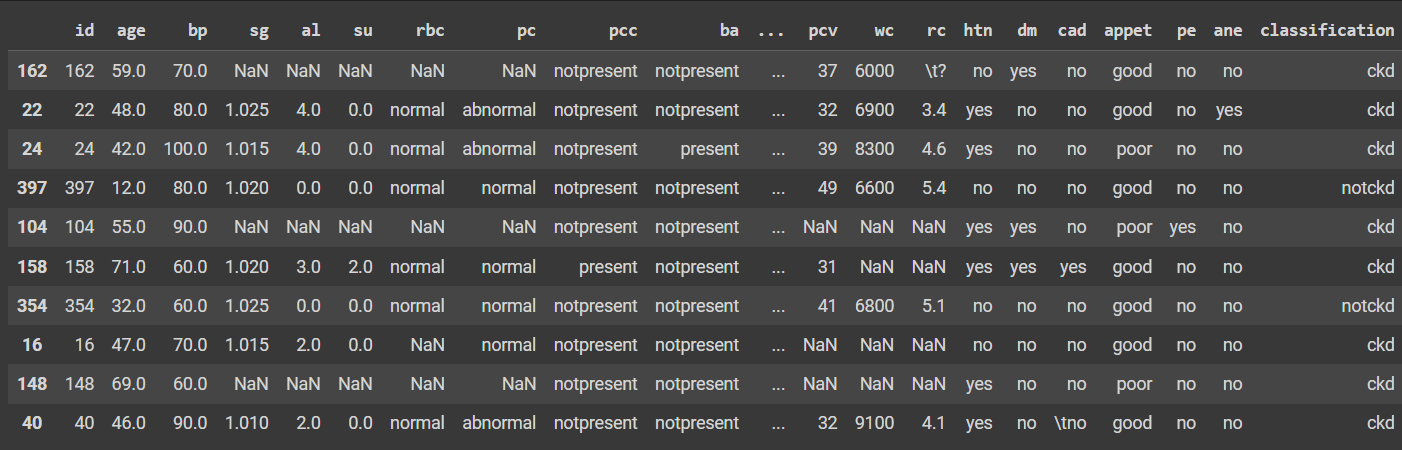
<https://colab.research.google.com/drive/1J5jvh3_DZN1asEBTz4OzxVH00YoK7YKp?usp=sharing>

**Results and Discussion**

**a. Description of the Proposed Datasets**

The dataset used includes 400 patient records with 25 features, sourced from the UCI Machine Learning Repository.

|  |  |
| --- | --- |
| Information | Details |
| Name of the dataset | Chronic Kidney Disease Data Set |
| Number of samples recorded | 400 |
| Source of the data | UCI Machine Learning Repository |
| Number of features | 25 |
| Number of class labels | 2 (CKD or non-CKD) |

****

**b. Details of the Hardware and Software Used**

* **Hardware:** Standard computing hardware (e.g., Intel i7 processor, 16GB RAM)
* **Software:** Python, Scikit-learn, XGBoost, CATBoost, ADABoost, Pandas, NumPy

**c. Literature Results in Terms of Evaluation Metrics**

Comparison metrics include accuracy

**d. Dataset Split-up**

* 30% of the data is used for training.
* 70% of the data is used for testing.

**e. Comparison of the Proposed Model's Performance**

**Model's Performance for XGBoost: -**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Model | Accuracy | | Bagging Tree Model | 96.2% | | Random Forest | 98.7% | | MLP | 97.5% | | SVM | 100% | | Random Tree | 97.5% | |

**Model's Performance for CATBoost: -**

|  |  |
| --- | --- |
| Model | Accuracy |
| Bagging Tree Model | 98.9% |
| Random Forest | 97.8% |
| MLP | 96.7% |
| SVM | 77.1% |
| Random Tree | 98.5% |

**f. Observations**

* The highest accuracy through CATBoost is Bagging Tree Model(98.9%)
* Feature selection using XGBoost, CATBoost, ADABoost improved model performance significantly.

**Conclusion**

The study successfully developed predictive models for CKD using machine learning algorithms, with Neural Networks, Random Forest, and Support Vector Machine showing the highest accuracy. Future work should focus on applying these models to other diseases and expanding the dataset for better generalizability.

**References**

1. V’asquez-Morales GR, Martinez-Monterrubio SM, Moreno-Ger

P, Recio-Garcia JA. Explainable prediction of chronic renal dis

ease in the colombian population using neural networks and case

based reasoning. IEEE Access. 2019;7:152900–10.

2. Sinha P, Sinha P. Comparative study of chronic kidney disease pre

diction using knn and svm. Int J Eng Res Technol. 2015;4:608–12.

3.Khan B, Naseem R, Muhammad F, Abbas G, Kim S. An empiri

cal evaluation of machine learning techniques for chronic kidney

disease prophecy. IEEE Access. 2020;8:55012–22.

4.Hosseinzadeh M, Koohpayehzadeh J, Bali AO, Asghari P, Souri

A, Mazaherinezhad A, Bohlouli M, Rawassizadeh R. A diagnostic

prediction model for chronic kidney disease in internet of things

platform. Multimedia Tool Appl. 2021;80(11):16933–50.

5.Gunarathne WHSD, Perera KDM, Kahandawaarachchi KADCP.

Performance evaluation on machine learning classification tech

niques for disease classification and forecasting through data

analytics for chronic kidney disease (ckd). In: 2017 IEEE 17th

international conference on bioinformatics and bioengineering

(BIBE). IEEE: UK; 2017. p. 291–6.

6.Alasker H, Alharkan S, Alharkan W, Zaki A, Riza LS. Detection

of kidney disease using various intelligent classifiers. In: 2017

3rd international conference on science in information technology

(ICSITech). IEEE; 2017. p. 681–4.

7. Abdullah AA, Hafidz SA, Khairunizam W. Performance compari

son of machine learning algorithms for classification of chronic

kidney disease (CKD). J Phys: Conf Ser. 2020;1529(5):052077.

8. Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattan

akit W, Suwannawach S, Ninchawee N. Predictive analytics for

chronic kidney disease using machine learning techniques. In:

2016 management and innovation technology in