A FIELD PROJECT REPORT

on

**“Chronic Kidney Disease Prediction using machine learning”**

**Submitted**

by

|  |  |
| --- | --- |
| 221FA04446  V.Jagadesh | 221FA04527  **B.Narendra kuamr** |
| 221FA04564 221FA04673  **K.Nithin Likitha** | |

**Under the guidance of**

**Ms.Sk.Sajida Sultana**

*Designation*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Chronic Kidney Disease Prediction using machine learning”** that is being submitted by 221FA04446 (V.jagadesh) ,221FA04527(B.Narendra kumar), (221FA04564(Nithin), (221FA04673(Likitha) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of **Ms. Sk.Sajida Sultana,**Assistant Professor, Department of CSE.

|  |  |  |
| --- | --- | --- |
| Guide name& Signature |  |  |
| Assistant/Associate/Professor, CSE | HOD, CSE | Dean, SOCI |



**DECLARATION**

We hereby declare that the Field Project entitled **“Chronic Kidney Disease Prediction using machine learning”** is being submitted by, 221FA04446(S. Jagadesh), 221FA04527(B. Narendra kumar), 221FA04564(K .Nithin) and 221FA04673(K .Likitha) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of **Ms .Sk .Sajida Sultana,**Assistant Professor, Department of CSE.

By

221FA04446-V. Jagadesh

221FA04527-B. Narendra kumar

221FA04564-K .Nithin

221FA04673-K .Likitha

Date:

## ABSTRACT

CKD is one of the most important burdens of disease occurring in the world, affecting millions. The disease is characterized by persistent and gradual loss of kidney function over time, if not discovered or treated, at such a stage that the progression leads to ESRD. It has become very crucial to detect CKD early on because patients are assured of better outcomes as it allows for timely intervention to slow the progress of the disease. Early stage of CKD lacks symptoms and with the conventional methods of diagnosis being expensive and simply not affordable to the majority, especially in low resource sites.

Machine learning has several advantages: it predicts CKD based on large amounts of clinical and patient data by analyzing many patterns and correlations between characteristics that do not emerge from traditional approaches. This aims to explore the possible capability of ML algorithms to predict CKD from existing medical data. Three machine learning techniques used in this case are Decision Trees (DT), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN), applied to the CKD dataset acquired from the UCI Machine Learning Repository.

The project highlights data preprocessing, feature selection, and model evaluation as integral steps for obtaining robust predictions. The key feature indicators of CKD were blood pressure, specific gravity, albumin, and age in this dataset. Special care has been taken regarding missing values, normalization, class imbalance of the dataset; these were challenging issues to deal with while developing the model.

The outcomes suggested that the model which captured the highest prediction accuracy was the Decision Tree model, closely followed by the Artificial Neural Network model. Although K-Nearest Neighbors is the simplest to implement, it showed a much lower degree of accuracy due to problems arising from class imbalance and scalability. These results suggest that, with proper feature selection and preprocessing, these ML algorithms can be helpful tools in predicting CKD and may allow for an earlier intervention for better prognosis of the patient.

Future work will include adding in extra functionality and further applying such models to larger, more heterogenous datasets. With the help of machine learning, the health care industry can develop low-cost, scalable solutions for the early detection of CKD and lighten the burden on healthcare systems and improve quality of life for individuals at risk for kidney failure.

**TABLE OF CONTENTS**

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Model Architecture for CKD Prediction |  |
| Figure 2. Confusion Matrix for Decision Tree Model Results |  |
| Figure 3. CKD Prediction System System Flow Diagram |  |
| Figure 4. Cost and Sustainability Analysis of the System  **LIST OF TABLES**  Table 1. Key Features in the CKD Dataset  Table 2. the Performance Metrics for Machine Learning Models |  |

**INTRODUCTION**

### INTRODUCTION

Chronic Kidney Disease, or simply CKD, is a slow and steady loss of renal function. The patient will suffer from CKD, which will result in the kidneys losing their capability to remove waste and excess fluids from the blood, eventually leading to renal failure with a necessity for dialysis or possibly transplantation. It is a significant global health problem with millions at risk and serious morbidity and mortality. Although much progress has been made in medical science, it is still quite challenging to diagnose CKD early because many patients suffer from it even with observable symptoms not in the early stages.

The current diagnosis processes for CKD are expensive, thus limited in access to all groups. Due to such constraints, researchers have begun paying more attention to alternative approaches that may predict the onset of CKD. EHRs and the increased utility of ML also present the potential to make more scalable and cost-effective models for predicting CKD. Machine learning algorithms may process enormous amounts of medical data to gain insights that may pass through the human eyes.

The project has applied several algorithms from the class of machine learning, namely Decision Trees, K-Nearest Neighbors, Artificial Neural Networks, in order to predict CKD basing on clinical data. The research is centered around understanding what features of the dataset are the most predictive features and how do the models rank in terms of accuracy, precision, and recall.

**CHAPTER-2**

**LITERATURE SURVEY**

**LITERATURE SURVEY**

#### Literature review

In the last few years, much research has been carried out using machine learning for CKD prediction. Researchers have used data either available in the hospitals themselves or the UCI Machine Learning Repository mainly to test a variety of techniques in ML. Decision Trees became one of the popular methods as it is simple and interpretable. Sharma et al. used a DT model for achieving a very high accuracy of 98.6%, whereas Chithra et al. used an ANN, which resulted in an accuracy of 95.2%.

It basically depends on proper feature selection and preprocessing for the success of machine learning models in predicting CKD. Among these features are age, blood pressure, specific gravity, and albumin levels, which play very important roles in improving model performance. Handling missing values and normalizing data are equally important procedures though, they really can make all the difference in enhancing the robustness of models. Preprocessing is all about having data cleaned and free of bias for avoiding wrong or inaccurate results.

Even so, there are still challenges such as class imbalance-the number of CKD instances is greatly fewer than the number of non-CKD cases-and the need for models to perform well across populations and scenarios.

#### 1.Motivation-

The motivation for this project arises from the ever-increasing global prevalence of Chronic Kidney Disease and the urging need to find early detection methods to mitigate its severe long-term consequences. CKD is a silent disease in which it often remains undiagnosed at its early stages, with obvious major morbidity and mortality worldwide. Millions of people are now diagnosed to be afflicted with CKD, and the World Health Organization has predicted that this will rank as the fifth cause of death by the year 2040. This results in extremely high treatment costs in terms of dialysis and kidney transplant, thus posing a tremendous burden on healthcare delivery systems, most significantly in less resource-intensive areas. Early detection and prevention, therefore, become crucial in reducing the costs borne by different healthcare systems and improving the outcome of the patients.

### PROPOSED SYSTEM

**3.1System Overview**

The system proposes to produce predictive values of CKD with the help of machine learning models. The system begins with preprocessing data for the cleanliness of data fit to building a model. By means of variable selection, critical features are selected as blood pressure, specific gravity, and albumin levels because these variables prove to be of significant values for prediction. After preprocessing, data input into multiple models of machine learning such as KNN, Decision Trees, and ANN for training and then evaluating.

**3.2Data Pre-processing**

The source of the data for this project was sourced from UCI Machine Learning Repository that has 400 records with 25 attributes. The preprocessing steps included dealing with missing values, scaling feature set, and also balancing up the dataset in order to prevent class imbalance.

**3.3Model Building**

For this project, three machine learning algorithms are selected:

KNN: It predicts CKD by comparing the test data with the 'k' nearest neighbors from the training data set.

DT: This model is a decision tree, in which it breaks down the dataset into smaller subsets based on the most important features.

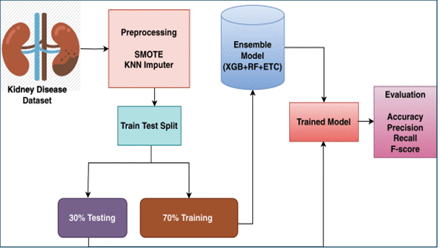
ANN: The model of the ANN mimics the structure of the human brain, in which it describes layers of neurons processing input data and producing predictions.

**3.4System Methodology**

The methodology of splitting the dataset into training and testing sets, where the models were then trained using the training set and evaluated by accuracy, precision, recall, and F1-score.

**3.5 Model Evaluation**

The critical performance metrics used for assessing the models were accuracy, precision, recall, and F1-score. As for accuracy, the decision tree performed better than the others, whereas ANN performs better when dealing with more complex data.



IMPLEMENTATION

**4.1 System Flow**

The system flow is designed to make the prediction process much easier, from data input to model evaluation. Flowchart \*\*Figure 4.1\*\* shows the two main steps of the system: pre-processing, training, and evaluation of the machine learning models.

**4.2 Cost and Sustainability Impact**

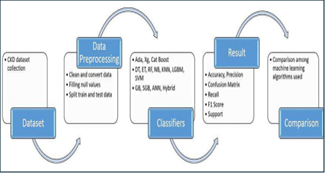
The system also becomes cost-effective as readily available medical data is used to predict the onset of early CKD, allowing health professionals to act on time and reduce the costs of managing CKD, such as dialysis and transplantation. The sustainability aspect is enhanced as the system uses electronic health records, thus reducing the use of expensive, invasive diagnostic tests.

The methodologies used in this project to predict Chronic Kidney Disease (CKD) are founded upon the established principles of machine learning and data analysis. The next sub-sections outline the major methodologies used in the system:

**5.Methodologies**

**5.1 Data Collection**

The dataset used in this project was taken from the UCI Machine Learning Repository that contains 400 records with 25 relevant attributes for diagnosis of CKD. This set contains clinical features such as blood pressure, specific gravity, and albumin, with demographic variables.



**5.2Data Pre-processing**

Data pre-processing is essential to enhance input data quality. That comprises a few phases:

Handling Missing Values: Missing values were imputed by replacing missing values with the mean or median of the respective feature. There is no information loss.

Normalization: Features were scaled so that the independent variables are in the same scale, for example, when algorithms are sensitive to scale input data (e.g., K-Nearest Neighbors).

Balancing the Dataset: Class imbalance was taken care of using various techniques, including over-sample of the minority class or under-sample of the majority class for CKD and non-CKD case classifications, to achieve equal representation of both in the training dataset.

**5.3 Feature Selection**

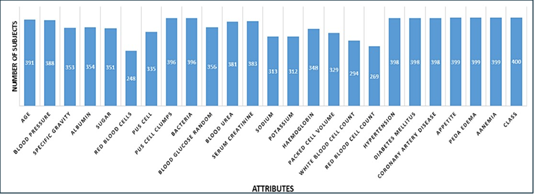
Feature selection is done to select those attributes which are the most relevant to the prediction of CKD. Some of the key features extracted are:

Blood pressure

Specific gravity

Albumin

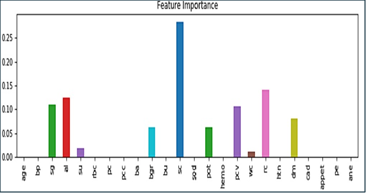
Age



According to the literature review, these characteristics have been selected since they were concerning CKD outcomes and very relevant to the earlier studies.

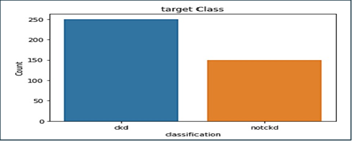
**5.4 Model Development**

Within this study, three machine learning models with the help of predictive models were applied; namely,Decision Trees (DT): This is a non-linear model where splitting of data into subsets occurs based upon the values of features so that a tree-like structure is formed. It is pretty convenient to interpret and friendly in terms of usage.



K-Nearest Neighbors (KNN): A distance-based algorithm that classifies a data point based on the majority class among its 'k' nearest neighbors in the feature space.

Artificial Neural Networks (ANN): A complex model that mimics the workings of the human brain using layers of interconnected neurons, suitable for capturing intricate patterns in data.



**5.5 Model Evaluation-**

The models were evaluated on numerous metrics:

Accuracy: The correctly predicted instances over the total instances.

Precision: The true positive predictions over the total predicted positives, which indicate how much correctness is with the class predictions of the positive class.

Recall: True positives over the actual positives, and is representative of how well the model can identify all the relevant instances in the dataset.

F1-Score: It is a harmonic mean of precision and recall. It offers equal weightage to both precision and recall.

A confusion matrix was used for the visualization of performance of models in making clear the difference between true and false predictions.

**RESULTS**

The promising results received from the models were that of decision tree models at an accuracy of 98.6%. The ANNN performed reasonably well, especially when it came to recognizing complex patterns in the data. In fact, this algorithm was weaker at handling class imbalance and less accurate than the other two models.

**CONCLUSION**

From the results of the project, it reveals that machine learning models, decision trees in general and artificial neural networks to a great extent can predict CKD. Including the salient features of the blood pressure and specific gravity significantly enhanced the performance of the models. The future work will be dedicated to applying these models to large datasets as well as the feature exploration, including genetic markers for enhancing prediction accuracy of CKD. Moreover, research is expected that would allow these models in the clinical setting in the application of them as the method of early diagnosis in both cases; this would improve patient results and decrease health care costs.

**REFERENCES**

1. Sharma, S., et al., "Performance-based evaluation of machine learning techniques for CKD diagnosis."

2. Chithra, P., et al., "Prediction of Chronic Kidney Disease: A Machine Learning Perspective." 3. UCI Machine Learning Repository, "Chronic Kidney Disease Dataset."

4. Akhter T, Islam MA, Islam S. Artificial neural network based covid-19 suspected area identification. J Eng Adv 2020;1:188–194.

5. Arora M, Sharma EA. Chronic kidney disease detection by analyzing medical datasets in weka. Int J Comput Mach Learn Algor New Adv Mach Learn 2016;3:19–48.

6. Banik S, Ghosh A. Prevalence of chronic kidney disease in Bangladesh: a systematic review and meta-analysis. Int Urol Nephrol 2021;53:713–718.

7. Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. 2016 Management and Innovation Technology International Conference (MITicon). IEEE; 2016.pp. MIT–80.

8. Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int Urol Nephrol 2016;48:2069– 2075

9. Chittora P, Chaurasia S, Chakrabarti P, et al. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access 2021;9:17312–17334.

10. Cueto-Manzano AM, Cortés-Sanabria L, Martínez-Ramírez HR, Rojas-Campos E, GómezNavarro B, Castillero-Manzano M. Prevalence of chronic kidney disease in an adult population. Arch Med Res 2014;45:507–513.

11. Dua D, Graff C. UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml 2017.

12. Eroğlu K, Palabaş T. The impact on the classification performance of the combined use of different classification methods and different ensemble algorithms in chronic kidney disease detection. 2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO). IEEE; 2016. p. 512–516.

13. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intel Learn Syst Appl 2017;9(01):1.

14. Gudeti B, Mishra S, Malik S, Fernandez TF, Tyagi AK, Kumari S. A novel approach to predict chronic kidney disease using machine learning algorithms. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE; 2020. p. 1630–1635.

15. Heung M, Chawla LS. Predicting progression to chronic kidney disease after recovery from acute kidney injury. Curr Opin Nephrol Hypertens 2012;21:628–634.