ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

PREDICTIVE MODELING OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING TECHNIQUES

VANGALA SWATHI [1], PICHIGUNTLA JAYAKUMAR [2], POTHUPALLI KUSHAL SAGAR [3], REDDY REDDY SATHVIK REDDY [4], ANANTHOJU SIDHARTHA CHARY [5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. Machine learning technique has become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. For this perspective, Chronic Kidney Disease prediction has been discussed in this article. Chronic Kidney Disease dataset has been taken from the UCI repository. Seven classifier algorithms have been applied in this research such as artificial neural network, C5.0, Chi square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2 and random tree. The important feature selection technique was also applied to the dataset. For each classifier, the results have been computed based on (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority oversampling technique with full features. From the results, it is marked that LSVM with penalty L2 is giving the highest accuracy of 98.86% in synthetic minority over-sampling technique with full features. Along with accuracy, precision, recall, F-measure, area under the curve and GINI coefficient have been computed and compared results of various algorithms have been shown in the graph. Least absolute shrinkage and selection operator regression

selected features with synthetic minority over-sampling technique gave the best after synthetic minority over-sampling technique with full features. In the synthetic minority over-sampling technique with least absolute shrinkage and selection operator selected features, again linear support vector machine gave the highest accuracy of 98.46%. Along with machine learning models one deep neural network has been applied on the same dataset and it has been noted that deep neural network achieved the highest accuracy of 99.6%.

Index Terms – Chronic Kidney Disease (CKD), machine learning, deep neural network (DNN), Linear Support Vector Machine (LSVM), Synthetic Minority Oversampling Technique (SMOTE), Least Absolute Shrinkage and Selection Operator (LASSO), feature selection, disease prediction, medical diagnosis, healthcare analytics, classification, UCI CKD dataset.

I. INTRODUCTION

Chronic kidney Disease (CKD) means your kidneys are damaged and not filtering your blood the way it should. The primary role of kidneys is to filter extra water and waste from your blood to produce urine and if the person has suffered from CKD, it means that wastes are collected in the body. This disease is chronic because of the damage gradually over a long period. It is flattering a common disease worldwide [1]. Due to CKD may have some health troubles. There are many causes for CKD like diabetes, high blood pressure, heart disease. Along with these critical diseases, CKD also depends on age and gender. If your kidney is not working, then you may notice one or more symptoms like abdominal pain, back pain, diarrhea, fever, nosebleeds, rash, vomiting. There are two main diseases of CKD: (i) diabetes and (ii) high blood pressure. So that controlling of these two diseases is the prevention of CKD. Usually, CKD does not give any sign till kidney is damaged badly. CKD is being increased rapidly as per the studies hospitalization cases increase 6.23 per cent per year but the global mortality rate remains fixed [4]. There are few diagnostic tests to check the condition of CKD: (i) estimated glomerular filtration rate(eGFR) (ii) urine test (iii) blood pressure. The doctor also asks for urine test for kidney functionality because kidneys make urine.

If the urine contains blood and protein [6], that means your kidney is not working properly. Doctor measures blood pressure as Blood pressure range shows how your heart is pumping blood. If eGFR value reaches less

than 15, that means the patient has end-stage kidney disease. At this point, there are only available treatments: (i) dialysis and (ii) kidney transplant. Patient's life after dialysis depends on such factors as age, gender, frequency and duration of dialysis, physical movement of the body and mental health [7]. If dialysis is not possible, the doctor has only one solution, i.e., kidney transplantation. However, it is extremely expensive [8]. Therefore, it is critical noteworthiness in early recognition, monitoring and handling of the disease. It is essential to predict the striding of CKD with appropriate accuracy due to its dynamic and secretive nature in the early stages and patient abnormality.

Medical treatment of CKD is prescribed by the stage. Anything other than this, it is very imperative to characterize the organization of the infection because it gives a few indications. It underpins the assurance of fundamental intercessions and medications. Medical treatment is a very significant application area of intellectual intelligent systems [10]. Afterwards, Data mining can play a big role to find out hidden information from the huge patient medical and treatment dataset that doctors frequently obtain from patients to get pieces of knowledge about the symptomatic data and to execute precise treatment plans. Data mining can be categorized as the method of extracting hidden information from a huge dataset. Data mining strategies are connected and utilized broadly in various contexts and areas. Using data mining methods, we may predict, classify, filter and cluster data. The objective states the algorithm processing of a training set containing a set of attributes and targets. Data mining is suitable to mining in data if the dataset is huge but we can also do it with the help of machine learning with a small dataset.

The machine learning can also find data analysis and pattern detection [9]. A variety of health dataset is present so machine learning algorithms are best fit to improve the accuracy of diagnosis prediction [11]. As healthcare electronic dataset grows rapidly, machine learning algorithms are becoming more common in healthcare. Qin et al. proposed data assertion and sample diagnosis achievable in CKD diagnosis. KNN is used for data assertion.

Six classifiers algorithms used for accuracy of diagnosis: logistic regression, random forest, support vector machine, K-nearest neighbor, naive Bayes classifier and feed-forward neural network. In these classifiers random forest gives better accuracy, i.e., 99.75%. Vasquez-Morales et al. [14]

developed a neural network model for risk prediction of chronic kidney disease development on 40000 instances dataset and their model accuracy was 95%.

II. RELATED WORK

A. *Predictive Analytics Of Chronic Kidney Disease By Using Machine Learning*

Predictive analytics for healthcare using machine learning is a challenged task to help doctors decide the exact treatments for saving lives. In this paper, we present machine learning techniques for predicting the chronic kidney disease using clinical data. Four machine learning methods are explored including K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers. These predictive models are constructed from chronic kidney disease dataset and the performance of these models are compared together in order to select the best classifier for predicting the chronic kidney disease.

B. *Comparative Analysis of Machine Learning Methods to Detect Chronic Kidney Disease*

Chronic kidney disease (CKD) is a major burden on the healthcare system because of its increasing prevalence, high risk of progression to end-stage renal disease, and poor morbidity and mortality prognosis. It is rapidly becoming a global health crisis. Unhealthy dietary habits and insufficient water consumption are significant contributors to this disease. Without kidneys, a person can only live for 18 days on average, requiring kidney transplantation and dialysis. It is critical to have reliable techniques at predicting CKD in its early stages. Machine learning (ML) techniques are excellent in predicting CKD. The current study offers a methodology for predicting CKD status using clinical data, which incorporates data preprocessing, a technique for managing missing values, data aggregation, and feature extraction. A number of physiological variables, as well as ML techniques such as logistic regression (LR), decision tree (DT) classification, and -nearest neighbor (KNN), were used in this work to train three distinct models for reliable prediction. The LR classification method was found to be the most accurate in this role, with an accuracy of about 97 percent in this study. The dataset that was used in the creation of the technique was the CKD dataset, which was made available to the public. Compared to prior research, the accuracy rate of the models employed in this study is considerably greater, implying that they are more

trustworthy than the models used in previous studies as well. A large number of model comparisons have shown their resilience, and the scheme may be inferred from the study's results.

C. *Prediction of chronic kidney disease-a machine learning perspective*

Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. Machine learning technique has become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. For this perspective, Chronic Kidney Disease prediction has been discussed in this article. Chronic Kidney Disease dataset has been taken from the UCI repository. Seven classifier algorithms have been applied in this research such as artificial neural network, C5.0, Chi square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2 and random tree. The important feature selection technique was also applied to the dataset. For each classifier, the results have been computed based on (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority over-sampling technique with full features. From the results, it is marked that LSVM with penalty L2 is giving the highest accuracy of 98.86% in synthetic minority over-sampling technique with full features. Along with accuracy, precision, recall, F-measure, area under the curve and GINI coefficient have been computed and compared results of various algorithms have been shown in the graph. Least absolute shrinkage and selection operator regression selected features with synthetic minority over-sampling technique gave the best after synthetic minority over-sampling technique with full features. In the synthetic minority over-sampling technique with least absolute shrinkage and selection operator selected features, again linear support vector machine gave the highest accuracy of 98.46%.

D. *Diagnosing Chronic Kidney Disease Using Hybrid Machine Learning Techniques*

The chronic kidney failure is a serious health issue and if not detected and treated at the early stages, it can be very deadly. Hence the major objective of this paper is to develop a

reliable machine learning model which predicts the CKD with a high accuracy rate. The CKD data set is downloaded from the famous UCI ML repository but it suffers from a lot of missing values. To handle the missing values KNN Imputation is used. Feature selection is also performed with the help of information gain as the dataset is huge and hence the cost of modelling can be very high. Various other pre-processing steps like label encoding and Min-max normalization is performed to attain a clean dataset. After pre-processing, various ML algorithms like logistic regression, naïve bayes, artificial neural network and random forest are applied and their performances are compared with the help of various performance metrics. A hybrid of Random Forest and Adaboost algorithm is proposed and it achieves a better accuracy when compared to the other individual component models and hence it can be proved that the proposed hybrid model is much better and accurate in diagnosing CKD.

III. PROPOSED METHODOLOGY

The proposed system aims to implement a machine learning model for predicting Chronic Kidney Disease (CKD) using advanced machine learning techniques. The core objective is to leverage a machine learning algorithm to predict CKD based on a dataset obtained from the UCI repository. The dataset categorizes individuals into two groups: CKD (represented by 1) and non-CKD (represented by 0). The system utilizes a machine learning model, which is trained on preprocessed and selected features from the dataset to predict CKD status. The performance of the trained model is evaluated using a validation set, ensuring high prediction accuracy. The model is then deployed as a web application, allowing users to input their demographic and clinical data and receive a prediction of their CKD status. The system's user-friendly interface displays the results clearly, providing reliable and efficient CKD predictions. This machine learning-based solution not only reduces the risk of human error compared to manual diagnosis but also improves healthcare by facilitating early detection and intervention of CKD, potentially leading to better patient outcomes. The system's flexibility allows for customization, incorporating additional features or data sources, enhancing its accuracy and reliability in CKD prediction.

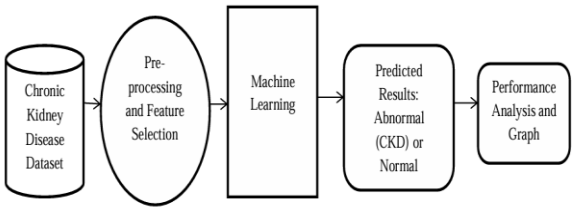

Fig. 1: System Overview

1. Data Transformation into Knowledge Graphs

Data Collection and Preprocessing: Begin with the collection of comprehensive employee data, including demographic details, job roles, performance metrics, and engagement scores. Preprocess this data to ensure consistency, handle missing values, and normalize numerical features. Graph Construction: Transform the pre-processed tabular data into a knowledge graph by mapping employees, departments, job roles, and other relevant entities as nodes, and defining relationships (e.g., reporting structures, project collaborations) as edges. This graph structure captures the complex interrelations within the organization, providing a richer dataset for analysis.

2. Hybrid Predictive Modelling

Enhanced Linear Support Vector Machine (L-SVM): Integrate the features extracted by the GCN into an enhanced L-SVM model. This hybrid approach leverages the linear separability of SVMs, augmented by the deep relational insights provided by the GCN, to improve turnover prediction accuracy. Benchmarking Against Traditional Models: Conduct a comparative analysis by training and testing the hybrid model against traditional machine learning algorithms such as Random Forest, Logistic Regression, and Gradient Boosting Machines. This benchmarking assesses the performance gains achieved through the integration of knowledge graphs and GCNs.

3. Explainable AI (XAI) Integration

Feature Importance Analysis: Implement XAI techniques, such as SHAP (Shapley Additive explanations) values, to interpret the contribution of each feature to the model's predictions. This analysis identifies the key factors influencing turnover, offering transparency and actionable insights. Visual Explanation Tools: Develop visual dashboards that illustrate the relationships between key factors and turnover risk, enabling HR professionals to easily understand and communicate the reasons behind model predictions.

4. Empirical Evaluation

Dataset Analysis: Apply the methodology to a real-world dataset of 1,470 IBM employees. Perform data splitting into training and testing sets, ensuring robust evaluation metrics such as accuracy, precision, recall, and F1-score. Case Study and Scenario Testing: Conduct case studies to validate the model's performance in

different organizational scenarios, such as changes in management structure or large-scale on boarding, to assess the model's adaptability and reliability.

5. Continuous Learning and Adaptation

Model Retraining and Update: Implement a system for continuous learning where the model is periodically retrained with new data to adapt to evolving organizational dynamics.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed machine learning and deep learning framework effectively predicts Chronic Kidney Disease (CKD) with high accuracy. Among the machine learning models, LSVM with L2 regularization achieved 98.86% accuracy using SMOTE with full features, while the Deep Neural Network (DNN) achieved the highest overall accuracy of 99.6%. The use of feature selection and SMOTE significantly improved classification performance by handling class imbalance.

Fig. 2: Home Page

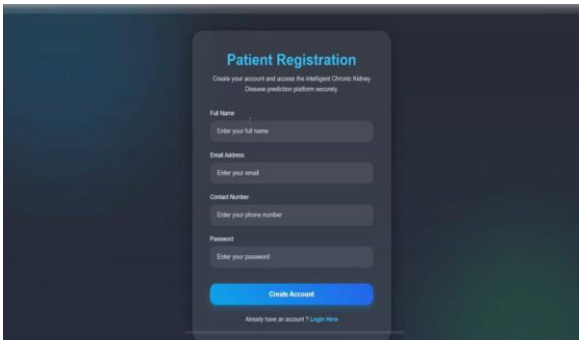

Fig. 3: Registration

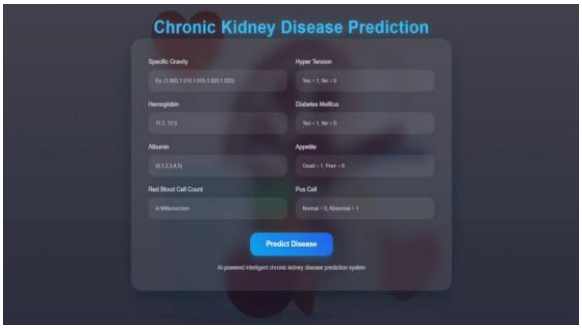

Fig. 4: Prediction

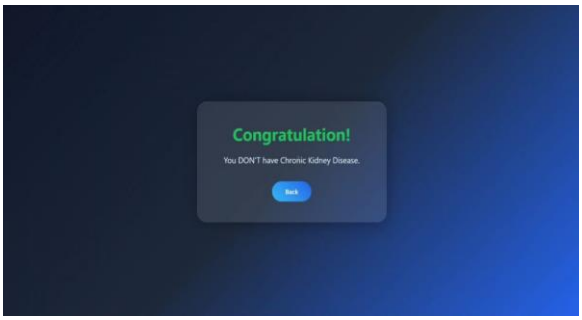

Fig. 5: Prediction Result

V. CONCLUSION

This article objects to predict Chronic Kidney Disease based on full features and important features of CKD dataset. For feature selection three different techniques have been applied: correlation based feature selection, Wrapper method and LASSO regression. In this perception, seven classifiers algorithm were applied viz. artificial neural network, C5.0, logistic regression, CHAID, linear support vector machine (LSVM), K-Nearest neighbors and random tree. For each classifier, the results were computed based on full features, selected features by CFS, selected features by Wrapper, selected features by LASSO regression, SMOTE with selected features by LASSO, SMOTE with full features. It was observed that LSVM achieved the highest accuracy of 98.86% in SMOTE with full features. All classifiers algorithms performed well on features selected by LASSO regression with SMOTE and without SMOTE. SMOTE with full features gave the best result for all 5 classifiers. In this research, a total of 7 classifiers were used. However, Logistic and KNN did not give suitable results and it was why they were not used in SMOTE. As per the result, it is concluded that SMOTE is a best technique for balancing a dataset. It is noted that SMOTE gave better results with selected features by LASSO regression as compare to without SMOTE on LASSO regression model. LSVM achieved the highest accuracy in all experiments as compared to other classifiers algorithms.

REFERENCES

[1] Q.-L. Zhang and D. Rothenbacher, ‘‘Prevalence of chronic kidney disease in population-based studies: Systematic review,’’ BMC Public Health, vol. 8, no. 1, p. 117, Dec. 2008.

[2] W. M. McClellan, D. G. Warnock, S. Judd, P. Muntner, R. Kewalramani, M. Cushman, L. A. McClure, B. B. Newsome, and G. Howard, ‘‘Albuminuria and racial disparities in the risk for ESRD,’’ J. Amer. Soc. Nephrol., vol. 22, no. 9, pp. 1721–1728, Aug. 2011.

[3] M. K. Haroun, ‘‘Risk factors for chronic kidney disease: A prospective study of 23,534 men and women in Washington County, Maryland,’’ J. Amer. Soc. Nephrol., vol. 14, no. 11, pp. 2934 2941, Nov. 2003.

[4] W. D. Souza, L. C. D. Abreu, L. G. D. SilvaI, and I. M. P. Bezerra, ‘‘Incidence of

chronic kidney disease hospitalisations and mortality in Espírito Santo between 1996 to 2017,'' Wisit Cheungpasitporn, Univ. Mississippi Medical Center, Rochester, MN, USA, Tech. Rep., 2019, doi: 10.1371/journal.pone.0224889.

[5] W. Mula-Abed, K. A. Rasadi, and D. Al-Riyami, ''Estimated glomerular filtration rate (eGFR): A serum creatinine-based test for the detection of chronic kidney disease and its impact on clinical practice,'' Oman Med. J., vol. 27, no. 4, pp. 339–340, 2012.

[6] A. S. Levey, D. Cattran, A. Friedman, W. G. Miller, J. Sedor, K. Tuttle, B. Kasiske, and T. Hostetter, ''Proteinuria as a surrogate outcome in CKD: Report of a scientific workshop sponsored by the national kidney foundation and the US food and drug administration,'' Amer. J. Kidney Diseases, vol. 54, no. 2, pp. 205–226, Aug. 2009.

[7] S. Gerogianni, ''Concerns of patients on dialysis: A research study,'' Health Sci. J., vol. 8, no. 4, pp. 423–437, 2014.

[8] J. R. Chapman, ''What are the key challenges we face in kidney transplantation today?'' Transplantation Res., vol. 2, no. S1, pp. 1–7, Nov. 2013.

[9] T. Xiuyi and G. Yuxia, ''Research on application of machine learning in data mining,'' in Proc. IOP Conf., Mater. Sci. Eng., 2018, doi: 10.1088/1757-899X/392/6/06220.

[10] B. Zupan, A. J. Halter, and M. Bohanec, ''Qualitative model approach to computer assisted reasoning in physiology,'' in Proc. Intell. Data Anal. Med. Pharmacol. (IDAMAP), Brighton, U.K., 2018, pp. 1–7.

AUTHORS Profile

Ms. VANGALA SWATHI has received her B.Tech from JNTUH, Hyderabad in 2015, M.Tech degree in Computer science from JNTUH, Hyderabad in 2017. She is dedicated to teaching field from the last 7 years. Her research area Deep Learning. At present she is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.

Mr. PICHIGUNTLA JAYAKUMAR pursuing B.Tech degree in Computer Science and Engineering at Guru

Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. POTHUPALLI KUSHAL SAGAR pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. REDDY REDDY SATHVIK REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. ANANTHOJU SIDHARTHA CHARY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.