e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 12, December 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712001

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

An Machine Learning Approach to Chronic Kidney Disease Detection

Loida Hermosure CpE, MIT, Rogelio Agustin Jr., Jomarie Amabao, Patrick De Leon,

Abegail Paladan, Zorayda Reazon, Jonathan Reazon

College of Information Technology, Northeastern College, Santiago City, Philippines

ABSTRACT: This study investigates the use of machine learning algorithms, specifically K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), for classifying kidney diseases based on a clinical dataset sourced from Kaggle. The dataset includes vital parameters such as blood pressure, specific gravity, albumin levels, sugar levels, hemoglobin, red and white blood cell counts, serum creatinine, and other clinical markers. Data preprocessing techniques, including imputation, normalization, and feature selection, were applied to ensure the dataset's quality and optimize model performance. The models were evaluated using accuracy, ROC-AUC scores, and confusion matrices to assess their classification capabilities. The Linear SVM model achieved a remarkable accuracy of 97.5% and an AUC score of 0.9987, demonstrating near-perfect discriminatory power and minimal false positives. In comparison, the Fine KNN model exhibited an accuracy of 95% with an AUC score of 0.9467, indicating robust performance but slightly lower precision in distinguishing between classes.

I.INTRODUCTION

A major and growing global health concern, chronic kidney disease (CKD) is becoming more widely acknowledged as a major cause of morbidity and mortality on a global scale. Due to the disease's covert course, many patients come with advanced renal insufficiency, which usually delays detection. A thorough analysis that covered the years 1990–2013 revealed a 90% increase in mortality from CKD, making it the 13th top cause of death worldwide. This frightening increase in deaths was brought to light by the report [1] Early-stage chronic kidney disease usually shows no symptoms. This is because the human body can usually adapt to a significant decline in kidney function. Until this point, kidney disease is frequently not detected unless a regular test for another condition, like a blood or urine test, identifies a possible problem. It may be prevented from developing into a more advanced form if it is identified early and treated with medication and regularly monitored with testing [2]. To diagnose renal illnesses, doctors typically use a mix of physical examinations and laboratory tests, including blood and urine analysis. To determine renal function and health, these tests measure albumin levels and glomerular filtration rate (GFR), respectively. The development of reliable and broadly applicable diagnostic models is essential in the current environment, which is marked by the availability of potential data sources [3]. Medical experts can make prompt and precise decisions with the help of these models [4]. One area of artificial intelligence called machine learning has the ability to detect CKD early, allowing for quick and effective treatments. By using past patient data to forecast future patient outcomes, this field has become crucial in the diagnosis of diseases. A variety of factors, including as diabetes, blood pressure, age, gender, smoking habits, creatinine levels, hypertension, and cholesterol, can be evaluated using machine learning algorithms for the prognosis of chronic kidney disease. Effective feature selection is essential to the performance of machine learning algorithms. By improving model accuracy, decreasing overfitting, accelerating training, streamlining interpretation, and avoiding data leaking, efficient feature selection expedites machine learning. altered a number of computing characterizations, such as back propagation networks (BPN), randomized subspace, kNN, DT, NB, and linear discriminant analysis (LDA) classifiers. To predict CKD and lower the mortality rate from CKD, the decision tree (DT) and NB characterisation methodologies were applied [9, 10].

II. RELATED LITERATURE

Two additional capabilities of machine learning are data analysis and pattern recognition. [5][6] Due to the great diversity of health datasets, machine learning algorithms are the most suitable approach for improving the precision of

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

diagnosis prediction. The rapid expansion of electronic healthcare datasets is directly contributing to the increasing prevalence of machine learning algorithms in the healthcare sector [7].

A study on the use of machine learning for the early diagnosis of chronic kidney disease was carried out by M.A. Islam et al. (2023) [2].

400 instances with 24 attributes—11 numerical and 13 categorical—were used in their work. Principal Component Analysis (PCA) was used to identify important features for CKD prediction following preprocessing. With 98.33% accuracy with the original data and 99.16% accuracy after applying PCA, the XgBoost classifier fared better than competing techniques. Prior to PCA, other classifiers also attained an accuracy of 98.33%. Three outfit measures and six classifiers were integrated in the instructions provided by Polat et al. [8].

Several classifiers were employed, including k-nearest neighbors (kNN), naïve Bayes (NB), support nector machine (SVM), preference tables, random forest (RF), and J48. The writers of Polat et al. used Apriori and the k-means algorithm to investigate several potential therapies for chronic renal illness.

An examination that makes use of SVM, DT, NB, and KNN. Using a dataset of 400 cases with 24 characteristics, Alsekait et al. (2023)6 created an ensemble deep learning model to predict CKD. Prior to feature selection using techniques like mutual information and Recursive Feature Elimination (RFE), the process comprised data pretreatment, which included label encoding and outlier detection. Using a Support Vector Machine (SVM) for meta-learning, the model combined RNN, LSTM, and GRU models in a stacked fashion. With an accuracy, precision, recall, and F1 score of about 99.69%, this model demonstrated strong performance characteristics [11]. Arif, M.S. et al. (2023) [12] created a machine learning model that used hyperparameter optimization, feature selection using the Boruta algorithm, and sophisticated preprocessing to forecast CKD. They used a novel sequential data scaling strategy that incorporated min-max scaling, z-standardization, and resilient scaling, together with iterative imputation for missing values. Using the k-Nearest Neighbors (KNN) method and grid-search CV for optimization, the model, which was tested on the UCI CKD dataset with 400 instances and 24 features, achieved a 100% accuracy rate. friend. S. (2022) [13] used a dataset of 400 cases with 24 features from the UCI machine learning library to create a model to predict CKD. Logistic logistic regression (LR), decision tree (DT), and support vector machine (SVM) classifiers were used in the study, and a bagging ensemble approach enhanced model performance. After using the bagging method, the Decision Tree (DT) classifier's accuracy rose from its peak of 95.92% to 97.23%.

To make informed clinical decisions regarding testing, treatment, and referral, it is crucial to have accurate information about the risk of nephropathy progression. This section focuses on recent advancements in chronic kidney disease (CKD) research. Our study explores the potential of various machine learning algorithms to facilitate the early diagnosis of CKD. While significant research has already been conducted in this area, we aim to enhance the approach by leveraging predictive modeling. Specifically, our methodology examines the relationships between data variables and target class characteristics. By incorporating predictive modeling, we can introduce more precise attribute measurements, enabling the development of robust prediction models through machine learning and predictive analytics.

III. METHODOLOGY

This study applies machine learning techniques, specifically K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), to classify kidney diseases based on a dataset obtained from Kaggle. The dataset includes key clinical parameters such as blood pressure (Bp), specific gravity (Sg), albumin levels (Al), sugar levels (Su), red blood cell count (Rbc), blood urea (Bu), serum creatinine (Sc), sodium (Sod), potassium (Pot), hemoglobin (Hemo), white blood cell count (Wbcc), red blood cell count (Rbcc), and the presence of hypertension (Htn). The methodology follows a structured process to ensure robust and reliable analysis.

The initial step involves data preprocessing to clean and prepare the dataset for analysis. Missing values are handled using appropriate imputation techniques, ensuring the dataset's integrity. Categorical variables, such as the presence or

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

absence of certain conditions, are converted into numerical formats using encoding methods like label encoding. To maintain uniformity, numerical variables are normalized using scaling techniques such as Min-Max or Standard Scaling, allowing the models to interpret the data more effectively. Additionally, correlation analysis or feature selection methods are applied to identify the most significant predictors of kidney disease, which optimizes the models' performance by reducing irrelevant or redundant information.

Once the data is preprocessed, the dataset is split into training and testing sets, typically with an 80-20 ratio. The training set is used to train the models, while the testing set evaluates their performance. For KNN, the model is trained by determining the optimal value of K through cross-validation and using Euclidean distance as the metric to measure the similarity between data points. For SVM, both linear and non-linear kernels, such as the radial basis function (RBF), are explored. Hyperparameters like the regularization parameter C and the kernel coefficient γ are tuned.

Model evaluation is conducted using a comprehensive set of metrics, including accuracy and ROC-AUC. These metrics provide insights into the models' classification performance, particularly their ability to handle imbalanced data. A confusion matrix is also generated to offer a detailed breakdown of correctly and incorrectly classified instances. The performance of KNN and SVM is compared to identify the superior model for this specific application, with a focus on their strengths and limitations in classifying kidney diseases.

Bp	Sg	Al	Su	Rbc	Bu	Sc	Sod	Pot	Hemo	Wbcc	Rbcc	Htn	Class
80	1.02	1	0	1	36	1.2	137.53	4.63	15.4	7800	5.2	1	1
50	1.02	4	0	1	18	0.8	137.53	4.63	11.3	6000	4.71	0	1
80	1.01	2	3	1	53	1.8	137.53	4.63	9.6	7500	4.71	0	1
70	1.005	4	0	1	56	3.8	111	2.5	11.2	6700	3.9	1	1
80	1.01	2	0	1	26	1.4	137.53	4.63	11.6	7300	4.6	0	1
90	1.015	3	0	1	25	1.1	142	3.2	12.2	7800	4.4	1	1
70	1.01	0	0	1	54	24	104	4	12.4	8406	4.71	0	1
76	1.015	2	4	1	31	1.1	137.53	4.63	12.4	6900	5	0	1
100	1.015	3	0	1	60	1.9	137.53	4.63	10.8	9600	4	1	1
90	1.02	2	0	0	107	7.2	114	3.7	9.5	12100	3.7	1	1
60	1.01	2	4	1	55	4	137.53	4.63	9.4	8406	4.71	1	1
70	1.01	3	0	0	60	2.7	131	4.2	10.8	4500	3.8	1	1
70	1.015	3	1	1	72	2.1	138	5.8	9.7	12200	3.4	1	1
70	1.02	1	0	1	86	4.6	135	3.4	9.8	8406	4.71	1	1
80	1.01	3	2	1	90	4.1	130	6.4	5.6	11000	2.6	1	1
80	1.015	3	0	1	162	9.6	141	4.9	7.6	3800	2.8	1	1
70	1.015	2	0	1	46	2.2	138	4.1	12.6	8406	4.71	0	1
80	1.02	1	0	1	87	5.2	139	3.7	12.1	8406	4.71	1	1
100	1.025	0	3	1	27	1.3	135	4.3	12.7	11400	4.3	1	1
60	1.015	1	0	1	31	1.6	137.53	4.63	10.3	5300	3.7	1	1
80	1.015	2	0	0	148	3.9	135	5.2	7.7	9200	3.2	1	1
90	1.02	1	0	1	180	76	4.5	4.63	10.9	6200	3.6	1	1
80	1.025	4	0	1	163	7.7	136	3.8	9.8	6900	3.4	1	1
70	1.01	0	0	1	57	3.07	137.53	4.63	12.53	8406	4.71	0	1
100	1.015	4	0	1	50	1.4	129	4	11.1	8300	4.6	1	1
60	1.025	0	0	1	75	1.9	141	5.2	9.9	8400	3.7	1	1

Figure (a) Sample Datasets

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712001

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

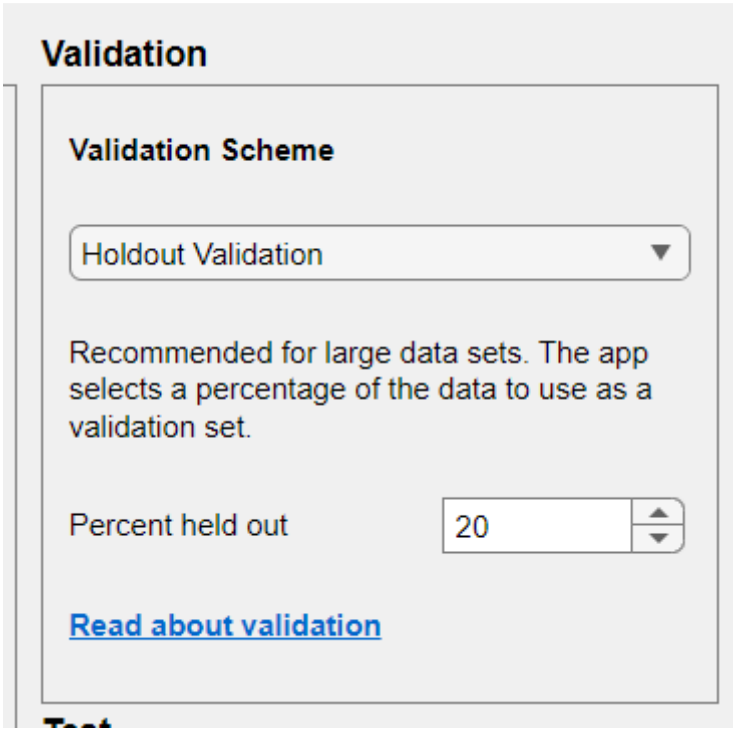

Figure (b) Pre Processing

IV. RESULTS AND DISCUSSION

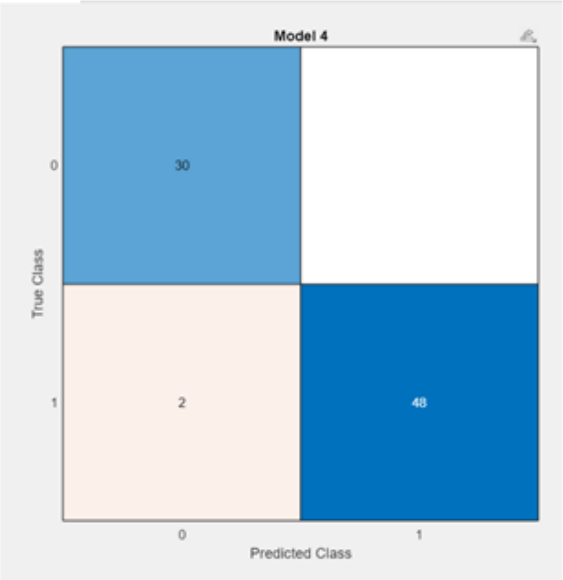

Figure (c) LINEAR SVM Confusion Matrix

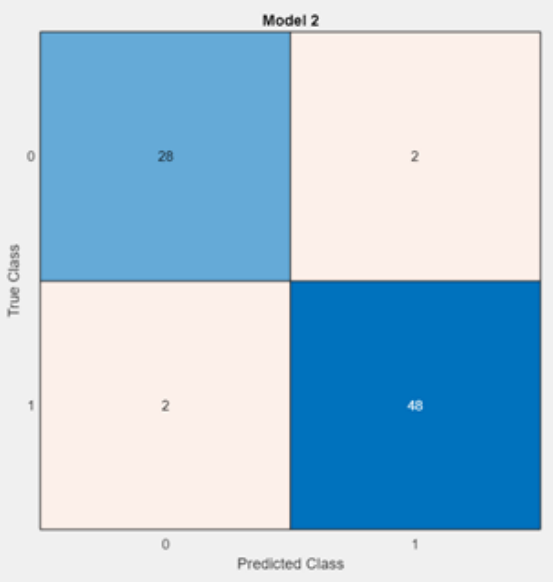

Figure (d) Fine KNN Confusion Matrix

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The performance of two machine learning models, Model 4 (Linear SVM) and Model 2 (Fine KNN), is compared using confusion matrices. These matrices summarize the models' ability to correctly classify instances in a binary classification task. Model 4, which employs a linear support vector machine (SVM), demonstrates strong performance, correctly identifying 48 instances of the positive class and 30 instances of the negative class. Notably, it produces no false positives, meaning it does not incorrectly classify any negative instances as positive. However, it does result in two false negatives, where positive instances are incorrectly classified as negative. Model 2, utilizing a fine K-nearest neighbors (KNN) approach, also performs well, correctly classifying 48 positive instances and 28 negative instances.

However, unlike Model 4, Model 2 generates two false positives, where negative instances are misclassified as positive. Similar to Model 4, it results in two false negatives. In comparing the two models, both achieve the same number of true positive and false negative classifications, indicating equal performance in identifying the positive class. However, Model 4's ability to avoid any false positives highlights its strength in distinguishing the negative class more accurately. On the other hand, Model 2 exhibits slightly reduced accuracy in classifying the negative class, as reflected in its false positive count.

To evaluate the performance of two machine learning classifiers, namely Fine K-Nearest Neighbor (KNN) and Linear Support Vector Machine (SVM), a methodology was designed to compare their Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values. The ROC curve serves as a graphical representation of the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at various threshold levels. A higher AUC value indicates better classifier performance, as it signifies the model’s ability to distinguish between classes effectively.

For the Fine KNN model, the ROC curve exhibits an AUC value of 0.9467, which reflects a strong ability to discriminate between the classes. The curve demonstrates a steep rise at the initial segment, indicating high sensitivity with a minimal increase in the false positive rate. However, the relatively lower AUC compared to the Linear SVM model suggests that there may be some limitations in its overall classification performance.

In contrast, the Linear SVM model achieves an AUC of 0.9987, which is nearly perfect and indicates exceptional discriminatory power. The ROC curve for this model sharply rises to the top left corner, suggesting that the classifier performs with very high sensitivity and low false positive rates. This superior performance underscores the Linear SVM’s capability to effectively learn the decision boundary between classes in the given dataset.

In contrast, the Linear SVM model achieves an AUC of 0.9987, which is nearly perfect and indicates exceptional discriminatory power. The ROC curve for this model sharply rises to the top left corner, suggesting that the classifier performs with very high sensitivity and low false positive rates. This superior performance underscores the Linear SVM’s capability to effectively learn the decision boundary between classes in the given dataset.

V. CONCLUSION

In conclusion, this study demonstrates the application of machine learning techniques, specifically K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), in the classification of kidney diseases using a Kaggle dataset that incorporates key clinical parameters. Through meticulous preprocessing, including imputation of missing values, normalization, and feature selection, the data was prepared to enhance the predictive performance of the models. The structured methodology, encompassing hyperparameter tuning and evaluation using metrics such as accuracy, confusion matrices, and ROC-AUC values, provided a comprehensive analysis of each model's strengths and limitations.

The results reveal that both KNN and SVM are effective classifiers for this task, but the Linear SVM model consistently outperforms Fine KNN in key areas. The Linear SVM achieved a near-perfect AUC value of 0.9987, showcasing exceptional discriminatory power and a superior ability to avoid false positives while maintaining high sensitivity. In comparison, the Fine KNN model, with an AUC of 0.9467, demonstrated strong performance but exhibited limitations in handling the negative class, as indicated by its false positive count.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," Journal of Big Data, vol. 9, no. 1, Nov. 2022, doi: https://doi.org/10.1186/s40537-022-00657-5.
[2] Md. A. Islam, Md. Z. H. Majumder, and Md. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," Journal of Pathology Informatics, p. 100189, Jan. 2023, doi: https://doi.org/10.1016/j.jpi.2023.100189.
[3] R. K. Halder et al., "ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application," Journal of Pathology Informatics, vol. 15, p. 100371, Feb. 2024, doi: 10.1016/j.jpi.2024.10037
[4] D. M. Alsekait et al., "Toward comprehensive chronic kidney disease prediction based on ensemble deep learning models," Applied Sciences, vol. 13, no. 6, p. 3937, Mar. 2023, doi: 10.3390/app13063937.
[5] T. Akhter, Md. A. Islam, and S. Islam, "Artificial Neural Network based COVID-19 Suspected Area Identification," Journal of Engineering Advancements, vol. 01, no. 04, pp. 188–194, Dec. 2020, doi: 10.38032/jea.2020.04.010.
[6] "Prediction of Chronic kidney Disease - a Machine learning perspective," IEEE Journals & Magazine | IEEE Xplore, 2021. https://ieeexplore.ieee.org/abstract/document/9333572
[7] M. A. Islam and Mosa. T. A. Shampa, "Convolutional Neural Network based Marine Cetaceans Detection around the Swatch of No Ground in the Bay of Bengal," 2022. https://www.semanticscholar.org/paper/Convolutional-Neural-Network-based-Marine-Cetaceans-Islam-Shampa/75316a75e5acefac730e671794cd00bd75ae83a2
[8] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," Journal of Medical Systems, vol. 41, no. 4, Feb. 2017, doi: 10.1007/s10916-017-0703-x.
[9] IEEE Conference Publication | IEEE Xplore, Aug. 01, 2015. https://ieeexplore.ieee.org/abstract/document/7275574
[10] International Center for Scientific Research and Studies, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors - DORAS." https://doras.dcu.ie/23782/
[11] D. M. Alsekait et al., "Toward comprehensive chronic kidney disease prediction based on ensemble deep learning models," Applied Sciences, vol. 13, no. 6, p. 3937, Mar. 2023, doi: 10.3390/app13063937.
[12] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the early detection of chronic kidney Disease: a robust machine learning model," Big Data and Cognitive Computing, vol. 7, no. 3, p. 144, Aug. 2023, doi: 10.3390/bdcc7030144.
[13] S. Pal, "Chronic kidney disease prediction using machine learning techniques," Deleted Journal, vol. 1, no. 1, pp. 534–540, Aug. 2022, doi: 10.1007/s44174-022-00027-y.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com