International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Enhancing Early Diagnosis of Chronic Kidney Disease Using Ensemble Model

***Mahalakshmi S*[1]*, Dr. Jerline Amutha*[2]**

Student, PG Department of Computer Science and Technology, Women's Christian College, Chennai, Tamil Nadu, India
Associate Professor, PG Department of Computer Science and Technology, Women's Christian College, Chennai, Tamil Nadu, India

ABSTRACT

Chronic Kidney Disease (CKD) is a prevalent medical condition that requires accurate early diagnosis to prevent severe complications, as delayed detection can lead to irreversible kidney damage and long-term health issues. Early intervention is crucial, as CKD often progresses silently with subtle symptoms, making it difficult to diagnose in its initial stages.

To address this challenge, this research focuses on developing a robust predictive model for CKD diagnosis using ensemble learning techniques. An ensemble model is constructed by combining Support Vector Machine (SVM), XGBoost, and Random Forest classifiers. The model leverages SVM's ability to handle high-dimensional data, XGBoost's proficiency in managing imbalanced datasets, and Random Forest's strong generalization capabilities. By harnessing the strengths of these machine learning algorithms, the ensemble model is able to capture complex patterns within the dataset. Trained and validated on a CKD dataset consisting of 600 instances, the ensemble model achieved an accuracy of 90.43%, highlighting its effectiveness and potential in enhancing early CKD diagnosis.

Keywords: Chronic Kidney Disease, ensemble learning, machine learning.

1. INTRODUCTION

Chronic Kidney Disease (CKD) is a significant and growing health challenge that affects a large portion of the global population. As a progressive condition, CKD often develops silently, with many individuals unaware of their deteriorating kidney function until it reaches a critical stage. This lack of awareness can lead to dire consequences, including irreversible kidney damage and heightened risk of cardiovascular diseases. Consequently, there is an urgent need for enhanced diagnostic tools that can facilitate early detection and intervention, ultimately improving health outcomes for affected individuals.

Traditional diagnostic methods, while foundational, may not always capture the nuances of CKD progression or adequately identify at-risk populations. In light of these limitations, innovative strategies utilizing advanced technologies are essential. Machine learning offers a transformative approach to healthcare diagnostics, enabling the analysis of complex datasets to uncover patterns that might escape conventional methods. Among these techniques, ensemble learning stands out by harnessing the strengths of multiple models, potentially leading to more accurate and reliable predictions.

This study seeks to develop a novel ensemble model that synergizes various machine learning algorithms to improve CKD diagnosis. By integrating multiple predictive strategies, the model aims to provide healthcare practitioners with enhanced decision-making capabilities, ultimately facilitating earlier intervention and better patient outcomes. Using a dataset of 600 instances, this research will explore how an ensemble approach can deliver robust predictions, paving the way for more effective management of chronic kidney conditions.

2. LITERATURE REVIEW

In recent years, machine learning techniques have been increasingly employed to enhance the accuracy and efficiency of CKD diagnosis. These approaches have the potential to recognize subtle patterns in clinical data that may go unnoticed in traditional diagnostic methods. This literature review explores various studies that have applied different machine learning models to predict CKD, with a focus on comparing the effectiveness of classifiers such as Support Vector Machines (SVM), Random Forest, XGBoost, and ensemble learning techniques. The collective findings highlight the importance of leveraging multiple algorithms and preprocessing strategies to improve diagnostic accuracy and early intervention.

Numerous studies have applied machine learning techniques to improve the prediction of CKD. For instance, [1] focused on SVM and Decision Tree, achieving an impressive 96.75% accuracy despite the challenges of small datasets, while [2] demonstrated the power of ensemble learning methods, achieving a 99% accuracy with a combination of classifiers. Both studies emphasize the importance of leveraging multiple algorithms to improve early CKD detection.

Similarly, [4] and [3] explored Random Forest in their approaches, with [4] achieving an accuracy of 98.89% and [3] reaching 99.75%. Both studies underscore the capability of Random Forest in handling large datasets with missing values, demonstrating its value in medical diagnosis. These findings are supported by [9], who employed Random Forest and XGBoost in an ensemble approach, further enhancing prediction accuracy for CKD.

On the other hand, [5] and [6] explored Gradient Boosting and other advanced classifiers. [5] implemented Gradient Boosting, yielding a remarkable accuracy of 99.80%, showcasing the potential of boosting techniques in early CKD detection. This aligns with [6], which emphasized exploring various classifiers for CKD prediction, highlighting the necessity for advanced machine learning techniques to improve diagnosis accuracy.

Lastly, [11] and [12] focused on refining datasets and key features in their models. [11] used AdaBoost to achieve the highest accuracy, emphasizing the importance of feature selection, while [12] performed thorough preprocessing of 25 variables, leading XGBoost to achieve an accuracy of 98.3%. These studies reflect the importance of data preparation in building reliable predictive models.

3. METHODOLOGY

The process begins with data collection, where relevant patient data is gathered from various sources. Following this, data preprocessing is performed to clean and prepare the dataset for analysis, ensuring that it is suitable for model training. Individual machine learning models, including Support Vector Machine (SVM), XGBoost, and Random Forest, are then implemented to assess their predictive capabilities. An ensemble model is constructed to combine the strengths of these algorithms, enhancing overall prediction accuracy. Finally, the model is rigorously evaluated to ensure its effectiveness in predicting CKD outcomes.

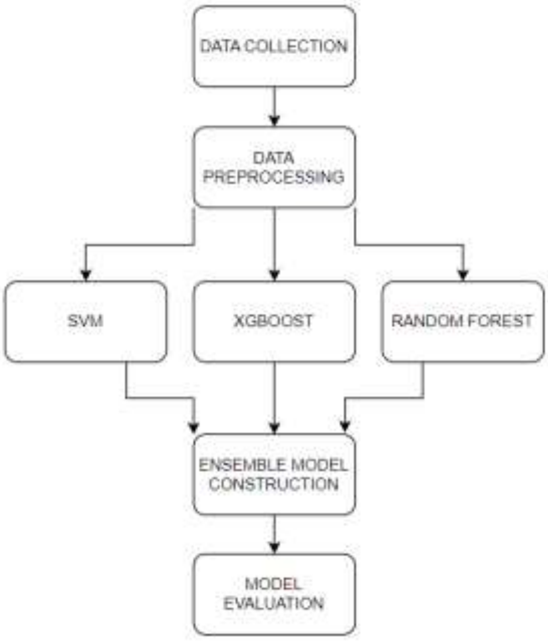


Figure 1: Workflow Diagram

3.1 DESCRIPTION OF DATASET

The dataset comprises a total of 600 instances, each containing 25 attributes, which are divided into two major categories: numerical and nominal data.

Numerical Attributes: These include measurable and continuous features such as:

- Age: The patient's age in years.
- Blood pressure (bp): The systolic blood pressure, typically measured in mmHg.
- Specific gravity (sg): A measure of the kidney's ability to concentrate urine, representing the density of urine relative to water.
- Albumin level (al): A measure of the presence of albumin in the urine, indicating kidney function.
- Blood glucose random (bgr): A measure of blood sugar levels at a random time of day.

- Blood urea (bu): The level of urea nitrogen in the blood, which helps assess kidney function.
- Serum creatinine (sc): The concentration of creatinine in the blood, an important marker for kidney performance.
- Sodium level (sod): The concentration of sodium ions in the blood.
- Potassium level (pot): The concentration of potassium ions in the blood.
- Hemoglobin level (hemo): The amount of hemoglobin in the blood, indicative of the oxygen-carrying capacity of red blood cells.
- Packed cell volume (pcv): The proportion of blood volume occupied by red blood cells, usually expressed as a percentage."
- White blood cell count (wc): The number of white blood cells per volume of blood, relevant to infection and immune response.
- Red blood cell count (rc): The number of red blood cells per volume of blood, important for evaluating anemia and overall health.

These numerical attributes represent measurable values critical for statistical analysis and predictive modeling.

Nominal Attributes: The categorical variables include:

- Red blood cells (rbc): This attribute indicates whether red blood cells are normal or abnormal in the urine.
- Pus cell (pc): Denotes the presence or absence of pus cells in the urine, which is important for diagnosing infections.
- Pus cell clumping (pcc): Indicates whether pus cells in the urine are clumped together, a potential sign of severe infection.
- Bacteria (ba): The presence or absence of bacteria in the urine, which can indicate a urinary tract infection.
- Hypertension (htn): A binary attribute indicating the presence (1) or absence (0) of hypertension.
- Diabetes mellitus (dm): Indicates whether the patient has been diagnosed with diabetes mellitus.
- Coronary artery disease (cad): Indicates the presence or absence of coronary artery disease.
- Appetite (appet): A measure of the patient's appetite, classified as good or poor.
- Pedal edema (pe): Refers to swelling in the lower limbs, a potential sign of kidney dysfunction.
- Anemia (ane): A binary indicator of whether the patient has anemia.
- Classification: This target variable indicates the presence or absence of Chronic Kidney Disease (CKD).

To ensure that categorical data can be effectively used in machine learning models, they will need to be encoded into numerical format. Together, the combination of numerical and nominal data enables a comprehensive analysis of CKD risk factors, contributing to the model's accuracy and reliability in prediction.

3.2 DATA PREPROCESSING

The preprocessing pipeline is critical to preparing the data for machine learning models. The following steps outline a comprehensive preprocessing strategy to ensure that the dataset is ready for predictive analysis:

- Handling Missing Values: Missing values in the numerical features are handled using the SimpleImputer class from sklearn, which replaces missing values with the mean of each feature. This strategy preserves the overall data structure while filling in the gaps, ensuring that no instances are dropped unnecessarily due to missing data.
- Scaling Features: The StandardScaler is applied to standardize the features. Standardization transforms the data such that each feature will have a mean of 0 and a standard deviation of 1. This step is especially crucial for machine learning algorithms that are sensitive to the scale of input data, such as support vector machines (SVM) and k-nearest neighbors (KNN).
- Outlier Detection: Outliers in the dataset, which can skew results and reduce the accuracy of the models, are detected and removed using IsolationForest. This algorithm is well-suited for identifying anomalies by isolating the data points that differ significantly from the majority. Removing outliers helps to improve the robustness of the predictive models.
- Class Imbalance: Since the dataset may have an imbalanced distribution between the classes (CKD vs non-CKD), the SMOTE (Synthetic Minority Over-sampling Technique) method is employed to generate synthetic samples for the minority class. This technique creates synthetic examples rather than simply duplicating existing instances, thus mitigating overfitting while balancing the class distribution. This step ensures that the models are not biased toward the majority class and are better equipped to predict both classes effectively.

This comprehensive preprocessing pipeline prepares the dataset by ensuring completeness, scaling, robustness against outliers, and balance in the class distribution, all of which enhance the model's ability to generalize and perform accurately on unseen data.

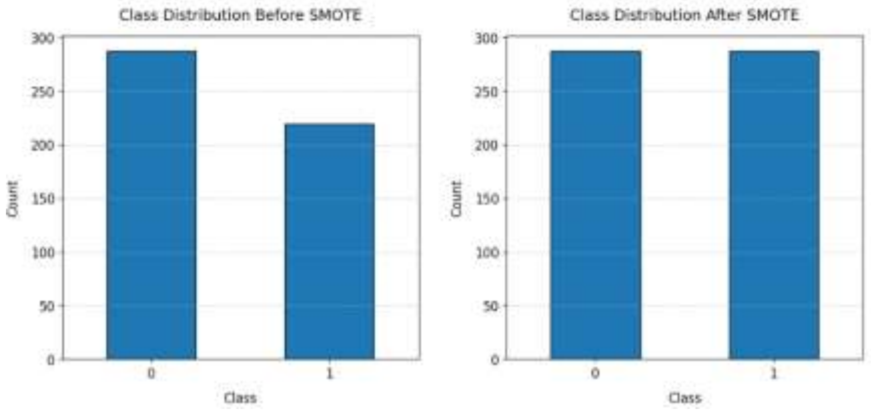


Figure 2: Class Distribution Before And After Applying Smote

3.3 MODEL BUILDING

In the pursuit of enhancing prediction accuracy for Chronic Kidney Disease (CKD), machine learning models have emerged as essential tools in the healthcare domain. These algorithms learn from historical data to identify patterns and make predictions, thereby providing valuable insights for clinical decision-making. Among the various techniques, the Support Vector Classifier (SVC), Random Forest, and XGBoost are particularly noteworthy. SVC is effective in high-dimensional spaces, while Random Forest enhances reliability through ensemble decision-making. XGBoost excels in capturing complex patterns within the data. By combining these models through an ensemble approach, we can leverage their individual strengths to achieve superior performance.

To enhance prediction accuracy for Chronic Kidney Disease (CKD), an ensemble approach is implemented using a StackingClassifier. This technique combines multiple base classifiers, each offering unique strengths to improve overall model performance. The model workflow begins by loading and splitting the dataset into training and testing subsets. To address the inherent class imbalance—typical in medical datasets—the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data. SMOTE generates synthetic samples for the minority class, balancing the class distribution and allowing the model to generalize better for underrepresented cases.

The stacking model integrates three base classifiers: Support Vector Classifier (SVC), Random Forest, and XGBoost. Each classifier plays a critical role in improving the prediction capability:

- Support Vector Classifier (SVC): SVC is known for its effectiveness in classification tasks, particularly in handling high-dimensional datasets where it constructs an optimal hyperplane for decision boundaries. It also provides the option for probabilistic outputs through Platt scaling, which helps improve interpretability in medical predictions.
- Random Forest: As an ensemble of decision trees, Random Forest introduces a high level of robustness by averaging the predictions of multiple trees. Its ability to handle both categorical and continuous variables makes it highly versatile, while its built-in feature importance ranking helps in understanding which factors most influence CKD predictions.
- XGBoost: This gradient-boosting algorithm is widely regarded for its efficiency and scalability in handling large datasets. XGBoost is particularly adept at identifying complex, non-linear relationships between features due to its iterative boosting process. Its regularization parameters also help to prevent overfitting, which is a crucial factor when dealing with medical datasets where overfitting can lead to misleading results.

In the stacking framework, a Random Forest is used as the final estimator. This meta-model aggregates the predictions of the base classifiers, combining their outputs to create a more comprehensive, accurate prediction. By leveraging the complementary strengths of these classifiers, the stacking approach enhances the model’s robustness and accuracy.

After applying SMOTE to the training set, the ensemble model is trained and evaluated on both the training and test sets. Key metrics such as accuracy, precision, recall, and F1 score are calculated to gauge its effectiveness in predicting CKD. These evaluation metrics are essential in assessing the performance of a predictive model. Accuracy measures the overall correctness of the model, while precision focuses on the proportion of correctly identified positive cases out of all predicted positives. Recall, or sensitivity, evaluates the model's ability to detect actual positive cases. The F1 score provides a balance between precision and recall, offering a single metric that accounts for both false positives and false negatives, making it particularly useful in imbalanced datasets like CKD prediction. This comprehensive approach ensures that the model not only performs well on the training data but also generalizes effectively to unseen test data. The resulting ensemble model becomes a highly reliable and accurate tool, contributing significantly to early detection and intervention for CKD, thereby aiding clinicians in making informed decisions.

4. RESULTS AND DISCUSSIONS

The ensemble model employed for predicting Chronic Kidney Disease (CKD) demonstrates remarkable predictive performance, achieving an accuracy of 90.43%. This result highlights the effectiveness of using a StackingClassifier, which integrates the strengths of multiple base models—Support Vector Classifier (SVC), Random Forest, and XGBoost—to enhance overall performance.

Individually, the classifiers performed admirably, with SVC achieving an accuracy of 86.96%, known for its ability to handle complex data patterns effectively. Random Forest followed closely with an accuracy of 87.83%, leveraging its ensemble of decision trees to provide robustness and high accuracy. XGBoost excelled as well, achieving an accuracy of 89.57%, owing to its gradient boosting techniques that adeptly capture intricate relationships within the data.

The ensemble approach effectively synthesizes the predictions of these diverse models, allowing for a more comprehensive understanding of the underlying data. By combining the unique strengths of SVC, Random Forest, and XGBoost, the stacking model improves upon the individual accuracies. This synergy enables the ensemble model to manage the complexities of CKD data more effectively, resulting in enhanced predictive capabilities.

The performance of the model is further confirmed by its confusion matrix results, which provide insights into its classification accuracy. For the training set, the model correctly identified 232 CKD-positive cases and 227 CKD-negative cases, with minimal errors. This shows excellent learning and accuracy during training. In the test set, the confusion matrix reveals that the model correctly predicted 50 CKD-positive cases and 46 CKD-negative cases, though it misclassified 11 CKD-negative cases as positive and 8 CKD-positive cases as negative. These metrics highlight the model's strong generalization performance and its ability to balance between detecting true cases and minimizing incorrect classifications.

Despite minor variations in performance among the base models, the ensemble's ability to outperform each individual classifier demonstrates the value of this integrated approach. The strong performance of the ensemble model reflects its reliability and effectiveness in CKD prediction, emphasizing its role in supporting healthcare professionals with informed decision-making for patient care.

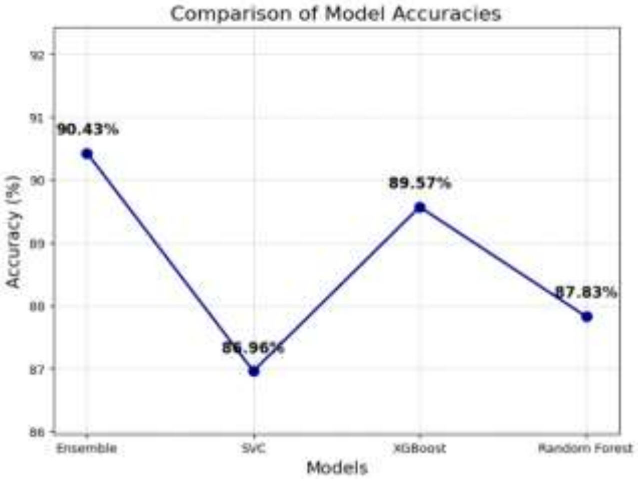


Figure 3: Models Accuracy Comparison

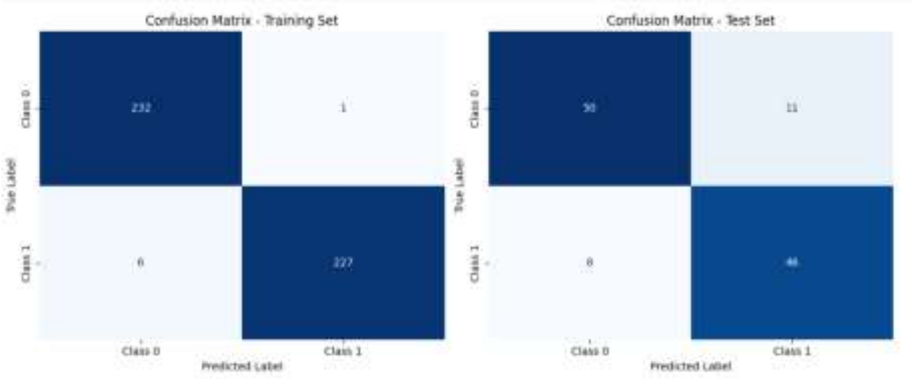


Figure 4: Confusion Matrix

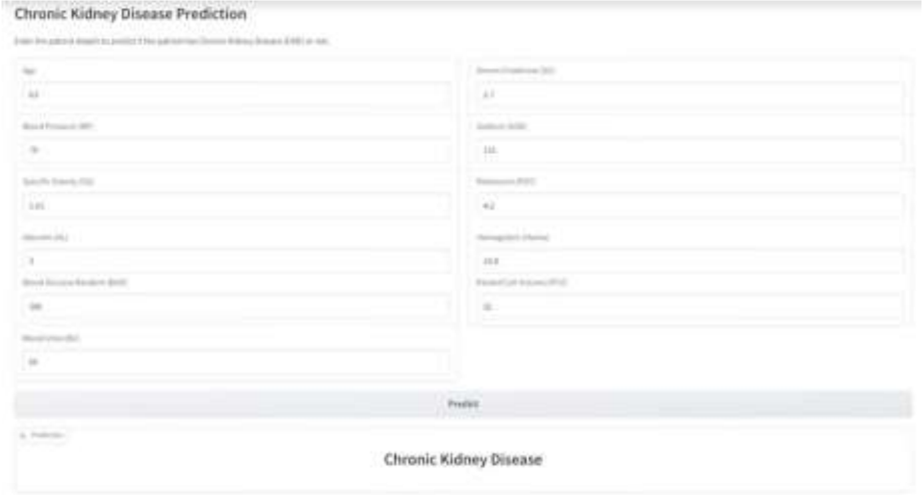
5. CONCLUSION

Chronic Kidney Disease (CKD) is a progressive condition that impairs kidney function over time, potentially leading to serious health complications if not diagnosed and managed early. The ensemble model offers a valuable advancement in the prediction of CKD, showcasing its potential to significantly enhance diagnostic accuracy. By combining the strengths of multiple algorithms—such as Support Vector Classifier (SVC), Random Forest, and XGBoost—the model effectively integrates diverse predictive insights, leading to more reliable and nuanced detection of CKD.

This comprehensive approach allows the model to capture complex patterns and subtle variations in the data that may be indicative of CKD. Its ability to leverage the unique contributions of each base classifier, along with a final estimator that consolidates these inputs, ensures a robust and well-rounded prediction capability. This can lead to earlier and more accurate identification of CKD, ultimately aiding in timely intervention and management.

Furthermore, the model's effectiveness in handling imbalanced datasets, through techniques like SMOTE, enhances its ability to accurately predict both positive and negative cases of CKD. This balanced approach is crucial for minimizing false positives and false negatives, which can have significant implications for patient outcomes.

Overall, the ensemble model holds the promise of improving CKD detection, offering a sophisticated tool for healthcare professionals to make more informed decisions and provide better patient care. Its advanced predictive capabilities can contribute to better disease management, earlier diagnosis, and ultimately, improved patient outcomes in the fight against CKD.



Chronic Kidney Disease Prediction

[illegible]

Figure 5: Prediction Using User Interface

References

[1] Tekale, S., Shingavi, P., Wandhekar, S., & Chatorikar, A. (2018). Prediction of Chronic Kidney Disease Using Machine Learning Algorithm. *International Journal of Computer Applications*.

[2] Hasan, K. M. Z., & Hasan, M. Z. (2019). Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease. *International Journal of Advanced Computer Science and Applications*.

[3] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. *Journal of Healthcare Engineering*.

[4] Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., & Hossain, S. (2020). Risk Factor Prediction of Chronic Kidney Disease Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*.

[5] Ghosh, P., Afrin, S., Mehedi Shamrat, F. M. J., Anjum, A. A., Shultana, S., & Khan, A. A. (2020). Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm. *Computational and Mathematical Methods in Medicine*.

[6] Revathy, S., Bharathi, B., Jeyanthi, P., & Ramesh, M. (2020). Chronic Kidney Disease Prediction Using Machine Learning Models. *Journal of Healthcare Engineering*.

[7] Gudeti, B., Mishra, S., Frederick Fernandez, T., Tyagi, A. K., Malik, S., & Kumari, S. (2020). A Novel Approach to Predict Chronic Kidney Disease Using Machine Learning Algorithms. *Journal of Healthcare Engineering*.

[8] Yashfi, S. Y., Islam, M. A., Pritilata, N., Sakib, N., Islam, T., Shahbaaz, M., & Pantho, S. S. (2020). Risk Prediction of Chronic Kidney Disease Using Machine Learning Algorithms. *Computational and Mathematical Methods in Medicine*.

[9] Wang, W., Chakraborty, G., & Chakraborty, B. (2020). Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm. *Journal of Healthcare Engineering*.

[10] Nishat, M. M., Faisal, F., Rahman, R., Nasrullah, S. M., Ahsan, R., Shikder, F., & Asif, M. A. (2021). A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms. *Computational and Mathematical Methods in Medicine*.

[11] Srikanth, V. (2023). Chronic Kidney Disease Prediction Using Machine Learning Algorithms. *International Journal of Computer Applications*.

[12] Islam, M. A., Hasan Majumder, M. Z., & Hussein, M. A. (2023). Chronic Kidney Disease Prediction Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*.