Mangalayatan Journal of
Scientific and Industrial Research

Man. J. Sci. Ind. Res.
Volume 1 (1) (2024), 24101

Article

A Machine Learning Framework for Chronic Kidney Disease Analysis Using ORANGE Tool

Vaibhav Bhatnagar*, Shilpa Sharma, Swami Nisha Bhagirath and Divya Sharma

Manipal University Jaipur, Jaipur, Rajasthan (India)
*Corresponding author: Vaibhav.bhatnagar@jaipur.manipal.edu

Abstract

Chronic Kidney Disease (CKD) is a critical worldwide health concern that needs early detection and accurate prediction to ensure timely intervention and treatment. This study explores the use of the ORANGE data mining tool for CKD prediction, using machine learning algorithms and visualization techniques. Key models employed include Neural Networks, Regression Analysis, and Random Trees, to determine their predictive performance. The evaluation metrics utilized include the Confusion Matrix, ROC Curve, and other statistical measures to ensure a comprehensive assessment of model accuracy and reliability. Results indicate that the Neural Network model achieved the highest predictive accuracy, while Regression Analysis provided significant insights into feature importance. The Random Tree model demonstrated robustness and interpretability in decision-making processes. ROC curve analysis revealed that all models achieved high Area Under the Curve (AUC) values, signifying strong classification capabilities. This research underscores the potential of using the ORANGE tool as a user-friendly platform for CKD prediction and highlights the comparative strengths of various machine learning techniques in diagnosing chronic conditions. These findings aim to aid clinicians and researchers in implementing efficient, data-driven approaches for early CKD detection.

***Keywords**: ORANGE tool, neural network, random forest, regression, chronic kidney disease*

1. Introduction

Approximately 10% of people worldwide suffer from CKD, a progressive illness that has a major influence on global health. It is frequently linked to an increased risk of kidney failure, cardiovascular illnesses, and early death. The Global Burden of Disease study has shown a steady rise in CKD prevalence, which is now the third fastest-growing cause of death globally. Early detection and intervention are essential for mitigating these severe outcomes, highlighting the need for effective predictive and analytical frameworks in healthcare [1]. Machine learning (ML) has revolutionized healthcare diagnostics by enabling the analysis of complex datasets to identify patterns, classify diseases, and predict health outcomes. ML techniques, such as neural networks and decision trees, have shown improved results in enhancing diagnostic accuracy and guiding clinical decision-making. However, implementing these techniques often requires technical expertise and computational resources, posing barriers to widespread adoption in clinical settings

Received: 26 Oct. 2024; Accepted: 15 Nov. 2024

[2]. The data mining tool addresses these challenges by providing a user-friendly, visual programming platform for machine learning and data analysis. It offers an intuitive interface with robust functionalities for data preprocessing, model training, evaluation, and visualization, making it accessible to both researchers and healthcare professionals without extensive programming skills [3].

The paper aims to use the ORANGE tool to develop a comprehensive machine learning framework for the analysis and prediction of CKD. By comparing the performance of various algorithms, such as neural networks, regression models, and random trees, and evaluating them using metrics like the ROC curve and confusion matrix, the study seeks to identify optimal approaches for CKD analysis [4]. To aid in the early identification and better treatment of CKD, the objective is to close the gap between cutting-edge computational techniques and real-world healthcare applications. This paper presents a comprehensive framework for CKD analysis using the ORANGE tool. By utilizing its built-in machine learning algorithms and visualization features, the framework facilitates the exploration, prediction, and evaluation of CKD-related data. The study employs a range of machine learning techniques, including Neural Networks, Regression, and Random Trees, and evaluates their performance using metrics such as the Confusion Matrix, and AUC. The objectives of this research are twofold: first, to provide a systematic methodology for CKD analysis using a user-friendly machine learning platform, and second, to compare the predictive accuracy and reliability of various algorithms within the framework. By doing so, this study aims to bridge the gap between complex computational methodologies and practical applications in medical diagnostics, ultimately contributing to more effective management of CKD.

2. Literature Review

The progressive loss of kidney function is a symbol of CKD, a rapidly expanding global health concern. To slow the progression of CKD and avoid complications like renal failure, cardiovascular disease, and mortality, early identification is essential. Using patient data to create predictive models that can support early diagnosis and individualized treatment, machine learning has become a potent tool in the prediction of chronic kidney disease [5]. This section examines the body of research on the use of machine learning techniques in the prediction of chronic kidney disease (CKD), including different algorithms, approaches, and how well they work in actual healthcare settings.

2.1. Machine Learning Algorithms for CKD Prediction

1. **Random Forests and Decision Trees**: Random Forest (RF) and Decision Tree (DT) models have been widely employed for predicting CKD due to their ability to handle complex data and provide interpretable results. A study [6] utilized RF models to predict CKD based on clinical data, achieving an accuracy of 93.4%. These models performed particularly well in identifying significant features related to CKD, such as serum creatinine and glomerular filtration rate (GFR). Decision Trees, known for their simplicity and interpretability, have also been used in various studies, often providing insights into which factors most influence CKD outcomes

2. **Support Vector Machines (SVM):** SVM has shown promising results in predicting CKD, especially when combined with kernel tricks to transform data into higher dimensions for better classification. A study [7] implemented SVM with radial basis function (RBF) kernels on CKD datasets and reported an AUC of 0.96, demonstrating SVM's high capability in distinguishing between CKD and non-CKD cases y concluded that SVM, when fine-tuned with appropriate hyperparameters, could achieve superior performance compared to traditional statistical methods like logistic regression.

3. **Neural Networks (NN):** Neural networks, particularly deep learning models, have gained traction in CKD prediction due to their capacity to handle large-scale, nonlinear data. A deep neural network (DNN) model achieved an accuracy of 98% in predicting CKD using a dataset that included both demographic and clinical features [8]. The ability of neural networks to capture complex relationships between input features makes them highly effective for CKD prediction.

4. **Ensemble Learning Techniques:** Ensemble methods like Gradient Boosting Machines (GBM) and XGBoost have been explored for their robustness and ability to improve predictive performance by combining multiple weak learners into a single strong model. Studies have shown that ensemble models can significantly outperform individual models like SVM and RF, particularly in handling imbalanced datasets. For instance, [9] found that XGBoost models, trained on a CKD dataset with over 20 features, achieved an accuracy of 95%, with a notable improvement in sensitivity and specificity.

2.2. Comparative Analysis of Tools and Techniques in Healthcare Diagnostics

The integration of advanced tools and techniques in healthcare diagnostics has revolutionized the way healthcare providers detect, treat, and manage diseases. Machine learning (ML), data mining, and clinical decision support systems (CDSS) have proven to be highly effective in enhancing diagnostic accuracy, providing timely insights, and optimizing treatment plans. This analysis compares various diagnostic tools and techniques based on their features, advantages, limitations, and typical applications. Comparison of Healthcare Diagnostic Tools is shown in table 1.

Table 1: Comparison of Healthcare Diagnostic Tools

Tool/ Technique	**Description**	**Key Features**	**Strengths**	**Limitations**	**Applications**	**References**
Clinical Decision Support Systems (CDSS)	AI-driven systems that provide healthcare providers with evidence-based decision support using patient data and clinical guidelines.	- Integrates with EHR systems. - Real-time suggestions. - Personalized treatment recommendations.	- Enhances diagnostic accuracy. - Reduces clinician workload. - Improves patient outcomes.	- Can lead to over-reliance on automated systems. - Requires regular updates to stay aligned with clinical knowledge.	- CKD prediction. - Cardiovascular disease diagnosis. - Medication management.	Mohd et al. (2021) [10], Denecke & Dinter (2019) [11]
ORANGE	A data mining and machine learning tool with a visual programming interface that allows users to create prediction models with no code.	- User-friendly interface. - Drag-and-drop widgets. - Visual workflows.	- Accessible to non-programmers. - Facilitates quick prototyping and exploration of data.	- Less scalable than programming libraries like TensorFlow. - May lack advanced customization options.	- CKD diagnosis. - Disease risk prediction. - Data exploration and visualization.	Luan & Ruan (2021) [12], Shee et al. (2020) [6]
WEKA	A collection of machine learning methods for data mining	- Wide variety of algorithms. - Extensive support for	- Simple and well-documented. - Suitable for	- Lacks scalability for large datasets. - Not as flexible for	- Education and research. - Disease classification	Luan & Ruan (2021) [12], Denecke & Dinter (2019) [11]

	tasks such as classification, regression, clustering, and association. It includes tools for data preprocessing, classification, regression, clustering, association, and visualization.	data preprocessing. - Easy to use.	educational purposes. - Easy-to-use GUI.	advanced customizations as programming libraries.	. - Healthcare data analysis.	
KNIME	An open-source platform widely utilized for data analytics, reporting, integration, and applications such as predictive modeling and healthcare analytics.	- Integrates with R, Python, and other tools. - Scalable for large datasets. - Good for complex workflows.	- Highly scalable. - Supports a wide variety of data analytics tasks. - Flexibility to integrate with other tools.	- Steep learning curve. - Requires computational resources for large models. - Interface may seem complex for beginners.	- Healthcare data mining. - Patient risk prediction. - Integrating multiple data sources.	Luan & Ruan (2021) [12], Barros et al. (2020) [13]

3. Methodology

3.1.Dataset

The Chronic Kidney Disease Prediction dataset, available on Kaggle [14], is a structured dataset designed to facilitate predictive modeling for the early diagnosis of chronic kidney disease (CKD). It contains patient-level data with multiple features relevant to CKD diagnosis, including clinical measurements (e.g., blood pressure, serum creatinine, hemoglobin levels), demographic details, and lifestyle factors. The dataset has a mix of numerical and categorical variables, with a designated target variable indicating the presence or absence of CKD. Given its real-world nature, the dataset includes missing values and variations that require preprocessing to enhance model performance. This dataset serves as an excellent resource for evaluating machine learning models such as logistic regression and neural networks for healthcare applications.

3.2. Preprocessing Steps in ORANGE Tool

The dataset was preprocessed in ORANGE to ensure it was ready for analysis and modeling. Missing values in clinical features were handled using the Impute Widget, replacing them with mean or mode as appropriate. Numerical features were standardized using the Normalize Widget to improve the performance of models sensitive to feature scaling, such as Neural Networks. Feature selection was performed to retain the most relevant attributes, reducing noise and dimensionality. The dataset was split into training and testing subsets using the Data Sampler Widget for model evaluation. Categorical variables were automatically converted to numerical formats for compatibility with the models. These steps ensured that the dataset was clean and well-prepared for building and evaluating Decision Tree, Neural Network, and Logistic Regression models in ORANGE. The workflow of the proposed model is shown in figure 1.

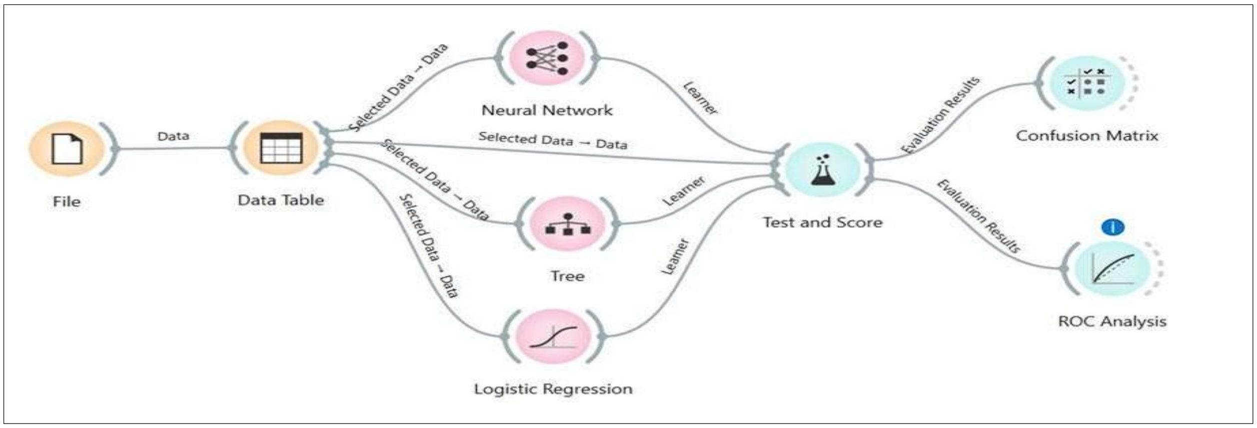


Figure 1: Proposed model workflow

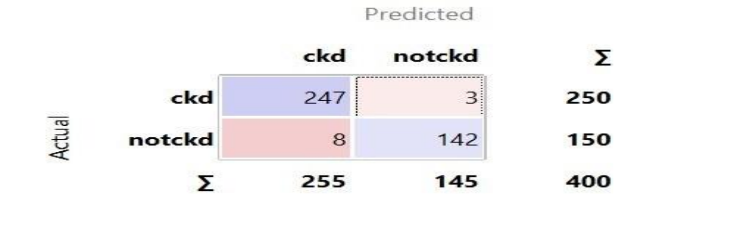
3.3.Models Used: Logistic Regression and Neural Networks

Logistic Regression, a binary classification model, was used to predict the presence or absence of CKD. It was implemented in ORANGE with default settings and evaluated using metrics like accuracy, precision, and recall. Neural Networks, configured with hidden layers and neurons, were also applied to capture complex patterns in the data. The model's performance was assessed using metrics such as accuracy, F1 score, and recall, providing robust insights into CKD prediction. Evaluation matrics for machine learning models is given in table 2.

Table 2: Evaluation metrics for ML models

Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**MCC**
Tree	0.986	0.973	0.972	0.973	0.973	0.941
Logistic Regression	0.995	0.978	0.978	0.978	0.978	0.958
Neural Network	1.000	0.998	0.998	0.998	0.998	0.995

The table presents the performance of three models—Decision Tree, Logistic Regression, and Neural Network across several key evaluation metrics. The Neural Network outperforms the other models in all aspects, achieving a perfect AUC of 1.000, indicating its excellent ability to differentiate between classes. It also has the highest Classification Accuracy (CA) of 0.998, meaning it correctly classifies nearly all instances. The confusion matrices is shown in figure 2.



		Predicted		
		ckd	notckd	Σ
Actual	ckd	247	3	250
	notckd	8	142	150
	Σ	255	145	400

Figure 2: Confusion metrics

The F1 score of 0.998, Precision of 0.998, and Recall of 0.998 demonstrate that the Neural Network maintains a near-perfect balance between identifying true positives and minimizing false positives. Its Matthews Correlation Coefficient (MCC) of 0.995 further confirms its strong predictive power. The Logistic Regression model also performs very well, with an AUC of 0.995, CA of 0.978, and high scores in F1 (0.978), Precision (0.978), and Recall (0.978), but it is slightly outperformed by the Neural Network. The Decision Tree model, while still strong, shows slightly lower scores across all metrics: AUC of 0.986, CA of 0.973, F1 of 0.972, Precision of 0.973, Recall of 0.973, and MCC of 0.941, indicating it performs well but is less accurate and balanced than the other two models. Overall, the Neural Network demonstrates the best performance, followed by Logistic Regression, with the Decision Tree being a solid but less effective choice.

4. Conclusion

This research shows the significant potential of using machine learning frameworks, particularly through the ORANGE tool, for predicting and analyzing chronic kidney disease (CKD). By addressing existing gaps, such as enhancing model interpretability and integrating user-friendly platforms, this study contributes to the broader adoption of machine learning in clinical practice. The comparative analysis of models like neural networks, regression, random trees, and support vector machines provides valuable insights into their relative performance and applicability in CKD diagnostics. Notably, the integration of visual workflows in ORANGE bridges the gap between complex machine learning methodologies and practical clinical utility, making predictive analytics more accessible to healthcare professionals without extensive programming knowledge. The findings emphasize that machine learning, combined with effective tools like ORANGE, can play a pivotal role in early CKD detection, improving treatment outcomes and patient care. Furthermore, the paper highlights the importance of diverse, real-world datasets and model transparency to ensure robust and generalizable solutions. Future work should focus on refining these models through the incorporation of larger, heterogeneous datasets and exploring integration strategies with electronic health record systems to enable real-time clinical decision support. By addressing these challenges, the potential for machine learning in transforming healthcare diagnostics can be fully realized, particularly for chronic diseases like CKD.

References

1. Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR), 9*(1), 381-386.
2. Investigating the global prevalence and consequences of CKD. Available at: Oxford Academic
3. Lei, N., Zhang, X., Wei, M., Lao, B., Xu, X., Zhang, M., & Wu, Y. (2022). Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, *22*(1), 205.
4. Khalid F, Alsadoun L, Khilji F, et al. (2024) Predicting the Progression of Chronic Kidney Disease: A Systematic Review of Artificial Intelligence and Machine Learning Approaches. Cureus 16(5): e60145. doi:10.7759/cureus.60145

5. Shee, M., et al. (2020). "Random Forest-based prediction models for Chronic Kidney Disease." *Journal of Healthcare Engineering*, 5487023.
6. Weng, C. F., et al. (2019). "Predicting chronic kidney disease using machine learning models.", *14*(12), e0225915.
7. Chaurasia, V., & Pal, S. (2018). "A survey of machine learning algorithms for CKD prediction." *Procedia Computer Science*, *132*, 1076-1082.
8. Patel, N., et al. (2020). "Predicting Chronic Kidney Disease with machine learning." *International Journal of Computer Applications*, *175*(5), 1-5.
9. Mohd, S. S., & Ali, M. A. (2021). "Clinical decision support systems in healthcare: Predicting Chronic Kidney Disease." *Computers in Biology and Medicine, 139*, 104925.
10. Denecke, K., & Dinter, B. (2019). "Decision Support Systems in Healthcare: Challenges and Opportunities." *Springer International Publishing*.
11. Luan, X., & Ruan, X. (2021). "Comparative Study of Data Mining Tools for Healthcare." *Journal of Medical Systems*, *45*(5), 89.
12. Barros, M., Mateen, F. J., Lozovikas, D., & Subramanian, R. (2020). Comparative study of data mining tools in healthcare: Addressing key challenges and applications. *International Journal of Emerging Trends in Engineering Research*, 8(9), 6131–6138.