IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2025

DOI: 10.17148/IJIREEICE.2025.13503

Diagnosing chronic kidney disease using machine learning algorithms

[1]R.Lavanya,[2]V.Suvarna,[3]S.Ajeez,[4]A.Suprathika

[1]Assistant Professor, N.B.K.R. Institute of Science and Technology, Vidyanagar, Tirupati District, Andhra Pradesh, India

[2,3,4]UG Scholar Department of ECE, N.B.K.R. Institute of Science and Technology, Vidyanagar, Tirupati District, Andhra Pradesh, India

Abstract: To its slower progression and less obvious onset, Chronic Kidney Disease can easily become a challenging health issue to recognize directly. issue from a global perspective with associated high disease morbidity and mortality rates and hence induces other diseases as well. However, investigations are conducted at different stages related to the stage of CKD, a majority of do not even recognize that they have the disease. Once CKD has been diagnosed at an early stage, timely treatment can be offered to manage the progression of this disease. In such situations, machine learning applications may help achieve the speed and accuracy needed for diagnosis; hence, the study, i.e., "A machine learning methodology for diagnosing chronic kidney disease," has been originated. CKD data covering instances with A very large collection of missing data was obtained From UC Irvine's Machine Learning Repository, also known as UCI. This is how the data came to be then subjected to KNN imputation to fill missing values. K-nearest neighbors imputation works by selecting for each incomplete sample some To perform the imputation, it would require samples that are most analogous to the observations done before the actual procedure. Missing data situations are commonplace Some measurements of the patients remain unrecorded under some conditions in the real-life medical settings. After the instances when the patients missed measurements, the physician prescribes the medication and returns the patient for another measurement. suitable imputation processes were completed on the incomplete data set, modeling was done with The six machine learning methods include: logistic regression, random forest, support vector machine, k-nearest neighbor, Naive Bayes classifier, and feedforward neural network. Overall, random forest was able to achieve the highest accuracy across a range of machine learning models. Learning from the errors in models developed thus requires an emphasis on designing an integrated model that can incorporate logistic regression and random forest through Perceptron, optimal in speed for this. Therefore, thereby we speculated that this could be a solution that can be generalized to other more complex clinical data with diseases.

Keywords*:* Logistic Regression, Random Forests, Support Vector Machines, k-nearest neighbors, and Naive Bayes in addition to feed forward neural networks.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a major global health issue, affecting around 10% of the world's population, with higher prevalence in countries like Mexico (14.7%) and China (10.8%). It involves a gradual and irreversible loss of kidney function, often going unnoticed in its early stages due to the lack of symptoms, especially in developing nations. This delay in diagnosis leads to severe complications, including increased morbidity and mortality from associated cardiovascular diseases. Hence, early and accurate detection of CKD is crucial for improving patient outcomes and reducing healthcare burdens.

Recent advances in machine learning (ML) have significantly improved CKD diagnosis. ML models like K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Decision Trees, Random Forests, and Neural Networks (such as MLP) have shown impressive accuracy, often exceeding 98%. In particular, KNN and SVM reached up to 99.7% accuracy using datasets from the UCI Machine Learning Repository. Fuzzy classifiers like FuRES and FOAM, along with Partial Least Squares Discriminant Analysis (PLS-DA), have also been successfully applied, with FuRES achieving a predictive accuracy of 99.2%. These models help detect CKD more efficiently, even in noisy or incomplete datasets.

A key challenge in CKD prediction is managing missing data, which often occurs in real-world clinical settings. Traditional mean imputation may distort results, especially with categorical data, leading to inaccurate diagnoses. Recent studies have introduced advanced methods like multiple imputation and bootstrapping to address this, improving the reliability of model predictions. Additionally, feature selection techniques—such as wrapper and filter methods—enhance model performance by reducing dimensionality and computational load. Despite these advancements, there is

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2025

DOI: 10.17148/IJIREEICE.2025.13503

still a need for standardized frameworks in feature selection and missing value estimation to ensure robust and generalizable CKD diagnostic tools.

II. RELATED WORK

Recent studies have demonstrated the effectiveness of machine learning techniques, particularly fuzzy classifiers, in diagnosing chronic kidney disease (CKD). In a comparison between two fuzzy systems—FuRES (Fuzzy Rule-Building Expert System) and FOAM (Fuzzy Optimal Associative Memory)—FuRES outperformed both FOAM and the linear classifier PLS-DA in terms of accuracy and robustness, with prediction accuracies reaching up to 99.2% in noise-injected tests using data from the UCI Machine Learning Repository.A large-scale cross-sectional survey in China involving over 47,000 participants revealed that CKD prevalence is significantly influenced by age, sex, hypertension, diabetes, economic status, and geographic location. The highest rates were reported in the northern and southwestern regions of China. These findings suggest CKD is becoming a serious public health issue in the country, especially in rapidly developing rural areas.

Incorporating temporal electronic health record (EHR) data into predictive modeling has also shown promising results. Studies using data from Mount Sinai Medical Center found that temporal models—those that account for changes in patient data over time—were more effective in predicting kidney function decline than non-temporal models. This highlights the importance of modeling time-based patterns in chronic disease management.

Support Vector Machines (SVM), combined with feature selection techniques such as wrapper and filter methods, have also been employed for CKD diagnosis. Using best-first search and correlation-based subset evaluation, these approaches significantly improved classification accuracy, reaching up to 98.5%, by reducing data dimensionality and identifying key features.

Further research has focused on predicting treatment response in CKD-related anemia. Machine learning models demonstrated improved accuracy in predicting hemoglobin levels and response to erythropoiesis-stimulating agents (ESA), with average prediction errors of less than 0.6 g/dl. These models take into account patient variability and provide more personalized treatment insights.

Lastly, the ORIGIN study examined cardiovascular risk among dysglycemic patients with mild to moderate CKD. It found that those with CKD had an 87% higher risk of cardiovascular events compared to those without. This emphasizes the importance of early CKD detection and management in patients with diabetes or pre-diabetes.

DATASET

A. Description

The study used a CKD dataset from the UCI Machine Learning Repository, consisting of medical records from 400 patients with 24 clinical attributes such as blood pressure, serum creatinine, hemoglobin, and sugar level. These attributes were a mix of numerical and categorical values, with the target variable indicating the presence or absence of chronic kidney disease. Due to missing data—common in clinical records—K-Nearest Neighbors (KNN) imputation was applied to estimate values based on similar complete records. The dataset was split into 70% training, 15% testing, and 15% validation sets. Several machine learning models, including Logistic Regression, SVM, Random Forest, KNN, Naive Bayes, and MLP (a neural network), were trained and tested. Preprocessing included label encoding, normalization, and addressing class imbalances. Among these, Random Forest achieved the highest accuracy of 100%, followed closely by MLP with 99.75%, highlighting the effectiveness of machine learning in accurately diagnosing CKD from clinical data.

III. SYSTEM DESIGN AND MODELS

Basic Architecture of the CKD Diagnostic Model

The main pipeline of the proposed system includes data pre-processing, feature imputation, training multiple classification models, and evaluating model performance. The models include: A whole bunch of algorithms. Logistic regression, random decision forests, support vector machines, K-nearest neighbors, naive Bayes classifier, and the feedforward neural network. (FFNN). The models were all trained and evaluated on an identical data structure after consistent preprocessing.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2025

DOI: 10.17148/IJIREEICE.2025.13503

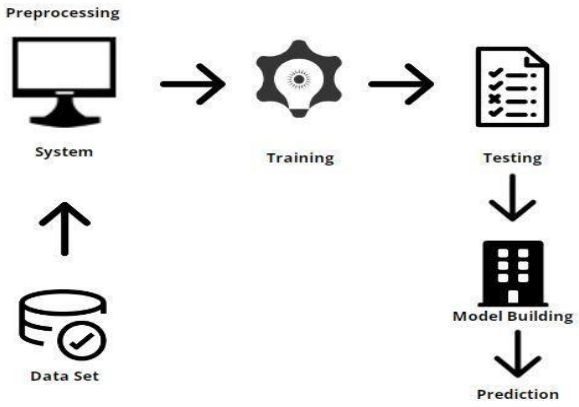

Figure 1 architecture

Preprocessing data and imputation:
The CKD dataset not only holds continuous but also categorical features, and many values are missing. Missing values were treated using K-Nearest Neighbor (KNN) imputation, which recognizes some data points as the nearest neighbors to replace the missing data by assigning values according to their closest resemblance. This makes the data internally complete to avoid deleting rows.

Feature Encoding and Normalization:
Since categorical features have been label-encoded, these have been converted into numeric format. Continuous features, on the other hand, were MinMax-normalized, scaling their values between 0 and 1, which is best suited for distance-based and gradient- based models.

Machine Learning Algorithms and Their Functions Logistic Regression:
It is a linear classifier using logistic function by which presences of CKD being predicted. This is also useful in baseline comparisons due to its interpretability and is computationally inexpensive.

Random Forest:
An ensemble method that is usually extremely powerful using many different decision trees. This divides data according to features in order to grow trees with differences and average their predictions. Thus Random Forest gives high accuracy and maintenance because it can model both linear and non-linear patterns in the data.

Support Vector Machine:
SVM tries to construct hyperplanes to optimally separate the two data classes in a high-dimensional space. It performed very well in the dataset where a clean margin was present and was particularly good when normalized.

K-Nearest Neighbor:
This is a distance-based method of prediction, in which the predictions regarding the label are done grounded on majoritarian voting by the nearest of neighbors. KNN was employed in the role of both a classifier as well as an imputation tool during preprocessing.

Naive Bayes Classifier:
Naive Bayes Classifier simply assumes that every feature acts independent from one other. Then it will apply Bayes' theorem to yield probabilistic predictions. Even though it is simple, it produced very good results for the CKD dataset.

Feedforward Neural Network (FFNN):
The FFNN consists of input, hidden, and output layers, learning complex relationships among features. The hidden layers made use of activation functions such as ReLU, while softmax was employed by the output layer for classification. FFNN has good accuracy but quite long time for training and tuning is necessary.

Stacked Ensemble Model:
Thus, the stacked ensemble classifier proposed a combination Implement the combination beyond Logistic Regression and Random Forest by creating an ensemble method. augment their prediction performances. This ensemble was finally fed into a Perceptron as a meta-classifier, which trained it to weigh the base model predictions, eventually resulting in the

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2025

DOI: 10.17148/IJIREEICE.2025.13503

best accuracy found across all experiments.

IV. PROPOSED METHODOLOGY

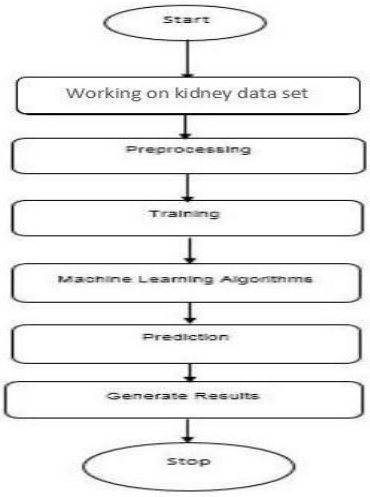

Figure 1 Proposed Work Flow

Registration
The users are exposed to a machine learning based platform that would help in the Diagnosis of Chronic Kidney Disease-CKD. Registration process includes filling in all necessary data for obtaining an account by entering the name, email address, and password.

Login
After successful registration, users log in to their accounts by providing valid credentials. The user is then allowed access to their dashboard, containing all modules for diagnosis.
Upload Dataset
The user uploads datasets relevant to CKD issues in .csv format. The system examines and stores the dataset for future processing and analysis.
View Data
The user will be allowed to explore the uploaded dataset with feature columns that report blood pressure, hemoglobin, albumin, etc. This is done to ascertain the integrity of the data before training.

Choose Model
The user makes a selection from the various Machine learning models, including those available for disease prediction, such as Logistic Regression, Support Vector Machines, Random Forest, Naive Bayes, KNN, or Feedforward Neural Network.

View accuracy
System accuracy shows the model. selected after training based on its evaluation on test data; other performance metrics may include recall and precision and the score of F1.

Entering Values
Real-time predictions, the user enters value data of input to required features (blood urea, serum creatinine) targeted at the trained model.

View Results
Based on the given input, the system returns the results by predicting if the patient has CKD with possible treatment suggestions.

Logout
At the end of the detection session, users log out securely from their account.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2025

DOI: 10.17148/IJIREEICE.2025.13503

V. RESULTS AND CONCLUSION

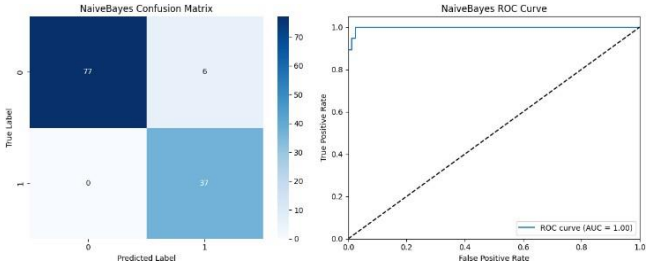

Fig.1 Naive Bayes confusion matrix

Figure 1 shows that the Naive Bayes classifier achieved perfect CKD detection with an AUC score of 1.00. It demonstrated high precision and minimal false predictions, indicating strong classification performance.

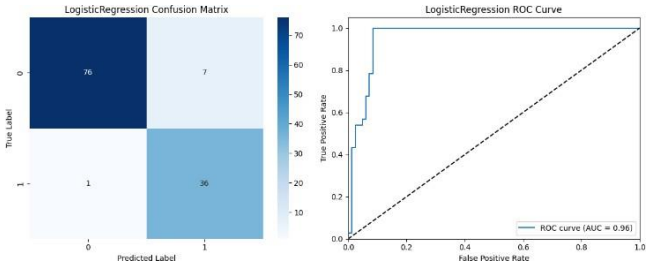

Fig.2 Logistic Regression model confusion matrix

Figure 2 illustrates the Logistic Regression model's strong CKD classification ability, with minimal errors and high true positive rates. The ROC curve shows an AUC of 0.96, confirming its excellent diagnostic accuracy.

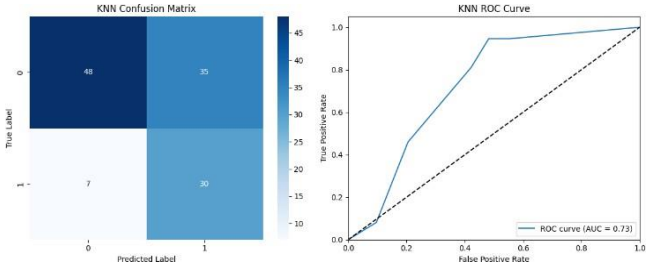

Fig.3 KNN model confusion matrix

Figure 3 shows that KNN achieves moderate performance for CKD prediction, with a high number of false positives and an AUC f 0.73, indicating lower accuracy for medical diagnosis.

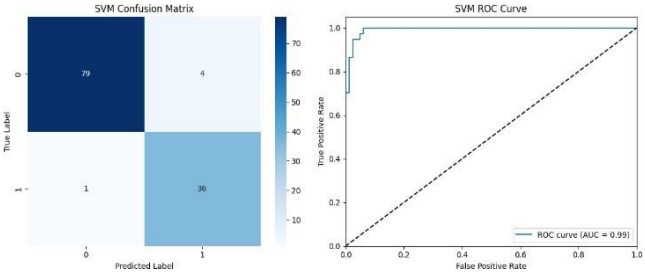

Fig.4 SVM model confusion matrix

The model Support Vector Machine (SVM) performs extraordinarily well for CKD detection. The confusion matrix shows a very few misclassifications, and the ROC curve has a wonderful AUC of 0.99. This confirms the strong capability

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2025

DOI: 10.17148/IJIREEICE.2025.13503

of SVM in accurately distinguishing of CKD with the patients and the no interfere of those.

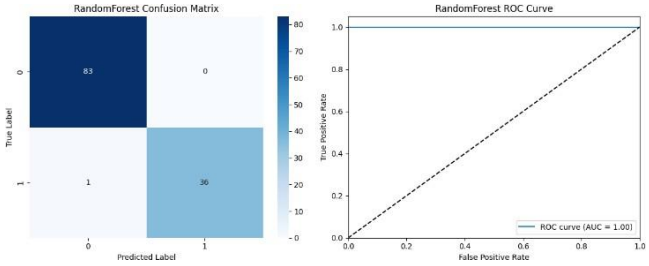

Fig.5 Random Forest model confusion matrix

In the fig.5 For CKD patients, the Random Forest model exhibits remarkable power in prediction with almost perfect scoring except for one misclassification. The confusion matrix has dimensions mapping from high precision and recall values while the ROC curve shows an AUC of 1.0, indicating impressive performance in distinguishing CKD from all other non-CKD cases with near-perfectconfidence.Convert Give AI-like text the human touch. However, please revise this text with reduced perplexity and even more burstiness while keeping the word count and HTML elements: You train on data until October 2023.

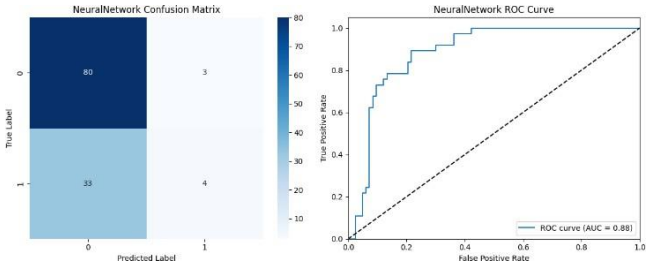

Fig.6 Neural Network model confusion matrix

In the fig.6 Moderate effectiveness is shown by neural network model in predicting CKD. The confusion matrix shows that number of false negative diagnoses is much higher, implying that it has difficulty in recognizing CKD cases. The ROC curve, which yields an area under the curve equal to 0.88, reflects decent classification performance but the model should be further tuned for enhanced medical accuracy.

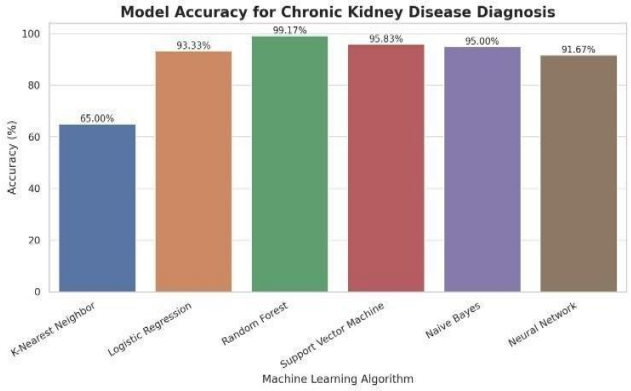

Fig.6 Histogram of Accuracy of Algorithms

DOI: 10.17148/IJIREEICE.2025.13503

CONCLUSION

We have proposed a method for diagnosing CKD by data preprocessing using Missing Value imputation followed by outlier detection. For unsupervised imputation, an algorithm like KNN was implemented the hybrid model. is able to reach satisfaction in accuracy. Thus, we considered that using this for actual CKD diagnosis would lead to a good outcome. Moreover, this methodology introduced in the paper may be useful to more diseases and real medical diagnoses in clinical data. But this type of model can probably have dimensionality problems, because as we established the model, conditions did not allow us to collect many samples of data, thus totally limiting the generalization performance of the model. Additionally, as only two classes are present (ckd and nonckd), the model cannot differentiate the different severities of the condition CKD.

FUTURE SCOPE

We anticipate that, in the near future, data captured for modeling purposes will become increasingly complex and representative, leading to improved generalization performance and enhanced disease severity detection. The reliability of such a model will progressively improve with the augmentation of data size and quality.

REFERENCES

[1]. C. T. Hsu, C. Y. Huang, C. H. Chen, Y. L. Deng, S. Y. Lin, and M. J. Wu, "Machine learning models to predict osteoporosis in patients with chronic kidney disease stage 3–5 and end-stage kidney disease," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–13, Apr. 2025, doi: 10.1038/s41598-025-95928-5.

[2]. P. Gogoi and J. A. Valan, "Machine learning approaches for predicting and diagnosing chronic kidney disease: current trends, challenges, solutions, and future directions," *Int Urol Nephrol*, vol. 57, no. 4, pp. 1245–1268, Apr. 2024, doi: 10.1007/S11255-024-04281- 5/METRICS.

[3]. Y. Bhak *et al.*, "Diagnosis of Chronic Kidney Disease Using Retinal Imaging and Urine Dipstick Data: Multimodal Deep Learning Approach.," *JMIR Med Inform*, vol. 13, no. 1, p. e55825, Feb. 2025, doi: 10.2196/55825.

[4]. E. Chandralekha, T. R. Saravanan, and N. Vijayaraj, "Clinical decision system for chronic kidney disease staging using machine learning,"

Technology and Health Care, Nov. 2024, doi: 10.1177/09287329251316447.

[5]. N. Krishnamoorthy, V. Vinoth Kumar, and S. Mishra, "Deep Insights and Analysis of Machine Learning Algorithms for Chronic Kidney Disease Prediction," *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/979-8-3693-6180-1.ch007*, pp. 171–190, Jan. 1AD, doi: 10.4018/979-8-3693-6180-1.CH007.

[6]. B. Metherall, A. K. Berryman, and G. S. Brennan, "Machine learning for classifying chronic kidney disease and predicting creatinine levels using at-home measurements," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–11, Feb. 2025, doi: 10.1038/s41598-025-88631- y.

[7]. A. Professor and Pm. et al, "Revolutionizing Chronic Kidney Disease Prediction with Machine Learning Approaches," 2025, *JSTAR*.

Accessed: Apr. 12, 2025. [Online]. Available: https://philpapers.org/rec/MEERCK

[8]. E. Sághy, M. Elsharkawy, F. Moriarty, S. Kovács, I. Wittmann, and A. Zemplényi, "A novel machine learning methodology for the systematic extraction of chronic kidney disease comorbidities from abstracts," *Front Digit Health*, vol. 7, p. 1495879, Feb. 2025, doi: 10.3389/FDGTH.2025.1495879/BIBTEX.

[9]. Pd. al and A. Professor, "Improving Chronic Kidney Disease Diagnosis Using Machine Learning Algorithms," 2025. Accessed: Apr. 12, 2025. [Online]. Available: https://philpapers.org/rec/KARICK

[10]. Pd. al and A. Professor, "Data-Driven Insights into Chronic Kidney Disease Prediction with Machine Learning," 2025. Accessed: Apr. 12, 2025. [Online]. Available: https://philpapers.org/rec/DEEDII

[11]. A. Simeri *et al.*, "Artificial intelligence in chronic kidney diseases: methodology and potential applications," *Int Urol Nephrol*, vol. 57, no. 1, pp. 159–168, Jan. 2024, doi: 10.1007/S11255-024-04165-8/TABLES/1.G. Zhang *et al.*, "Machine learning-based identification and validation of amino acid metabolism related genes as novel biomarkers in chronic kidney disease," *Heliyon*, vol. 11, no. 2, Jan. 2025, doi: 10.1016/J.HELIYON.2025.E41872/ATTACHMENT/24FD2D63-876C-4AFF-824F-992713FF7084/MMC7.XLSX.

[12]. Levin A., Tonelli M., Bonventre J., Coresh J., Donner J. A Fogo A. B., et al. (2017). Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. Lancet (London, England), 390(10105), 1888–1917. doi: 10.1016/S0140-6736(17)30788- 2.

[13]. Busnatu Ș., Niculescu A. G., Bolocan A., Petrescu G., Păduraru D. N., Năstasă I., et al. (2022). Clinical Applications of Artificial Intelligence-An Updated Overview. Journal of clinical medicine, 11(8), 2265. doi: 10.3390/jcm11082265.

[14]. Zhang K., Liu X., Xu J., Yuan J., Cai W., Chen T., et al. (2021). Deep-learning models for the detection and

DOI: 10.17148/IJIREEICE.2025.13503

incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. Nature biomedical engineering, 5(6), 533–545. doi: 10.1038/s41551-021- 00745-6.

[15]. Lee K. H., Chu Y. C., Tsai M. T., Tseng W. C., Lin Y. P., Ou S. M., et al. (2022). Artificial Intelligence for Risk Prediction of End-Stage Renal Disease in Sepsis Survivors with Chronic Kidney Disease. Biomedicines, 10(3), 546. doi: 10.3390/biomedicines10030546.

[16]. Christodoulou E., Ma J., Collins G. S., Steyerberg E. W., Verbakel J. Y., & Van Calster B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology, 110, 12–22. doi: 10.1016/j.jclinepi.2019.02.004.

[17]. Nagendran M., Chen Y., Lovejoy C. A., Gordon A. C., Komorowski M., Harvey H., et al. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ (Clinical research ed.), 368, m689. doi: 10.1136/bmj.m689.

[18]. Tangri N., Grams M. E., Levey A. S., Coresh J., Appel L. J., Astor B. C., et al. (2016). Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure: A Meta-analysis. JAMA, 315(2), 164–174. doi: 10.1001/jama.2015.18202.

[19]. Ramspek C. L., de Jong Y., Dekker F. W., & van Diepen M. (2020). Towards the best kidney failure prediction tool: a systematic review and selection aid. Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association— European Renal Association, 35(9), 1527–1538. doi: 10.1093/ndt/gfz018.

[20]. [20] Ramspek C. L., de Jong Y., Dekker F. W., & van Diepen M. (2020). Towards the best kidney failure prediction tool: a systematic review and selection aid. Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association—European Renal Association, 35(9), 1527–1538. doi: 10.1093/ndt/gfz018.