IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.12505

MACHINE LEARNING ALGORITHM FOR CHRONIC KIDNEY DISEASE PREDICTION

Sameena Firdaus[1], Sarfraj Alam[2], Somulapalli Navya[3], Julure Raviteja[4]

UG Scholars, Department of Computer Science and Engineering, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India[1,2,3]

Assistant Professor, Department of Computer Science and Engineering, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India[4]

Abstract: This study explores the use of machine learning to improve early detection of Chronic Kidney Disease (CKD). Using a data set from the UCI repository, seven classifiers and multiple feature selection techniques were evaluated. The Linear SVM with L2 regularization achieved 98.86% accuracy with SMOTE and full features, while a Deep Neural Network reached the highest accuracy of 99.6%. The results highlight the effectiveness of machine learning, especially deep learning, in enhancing CKD diagnosis.

Keywords: predictive modeling, SVM, logistic regression, neural network, random tree.

I. INTODUCTION

Chronic Kidney Disease (CKD) is a growing global health concern, affecting millions worldwide. Machine learning (ML) offers a promising approach for early CKD detection by analyzing patient data to identify subtle patterns indicating CKD risk. This project proposes developing a Java-based, ML-driven system for early CKD detection, utilizing algorithms like decision trees and support vector machines to predict CKD risk.

The system aims to be scalable, efficient, and integrable with existing healthcare systems, with objectives including system development, performance evaluation, and comparison of ML algorithms, ultimately potentially improving CKD detection, patient outcomes, and reducing healthcare costs.

In this study, several machine learning models—such as Linear Support Vector Machine (LSVM), Artificial Neural Network (ANN), and Logistic Regression—were applied to a CKD dataset. To boost the models' performance, feature selection techniques like LASSO and SMOTE were used.

II. EXISTING SYSTEMS

- The machine learning classifiers such as artificial neural network (ANN), C5.0, logistic regression, linear support vector machine (LSVM), K_x0002_nearest neighbors (KNN) and random tree were used for training the model.
- The procedure of this research including five stages:
  (i) dataset preprocessing,
  (ii) feature selection,
  (iii) classifier application,
  (iv) SMOTE
  (v) analyzing the performance of the classifier
- Along with machine learning models, a deep neural network was applied for comparing the result of machine learning models and deep neural network.
- Artificial Neural network classifier was used for this purpose. In this research the significance of two model were checked by statistic testing namely McNemar's test.

Disadvantages:

- Edge computing makes the environment more complex.
- The current edge computing architecture has the problems
- The edge node carries too many task requests

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.12505

III. PROPOSED SYSTEM

- This paper investigates different feature selection methods for predicting Chronic Kidney Disease (CKD) using machine learning techniques.
- The study tested several approaches, including using full features, correlation-based feature selection, Wrapper method, Least Absolute Shrinkage and Selection Operator (LASSO) regression, and Synthetic Minority Over-sampling Technique (SMOTE) with LASSO-selected features and full features.
- Among the machine learning models, the Linear Support Vector Machine (LSVM) with L2 penalty achieved the highest accuracy of 98.86% when using SMOTE with full features.
- Linear support vector machine (LSVM) is the modern particularly fast machine learning algorithm for solving multiclass classification problem for the large dataset based on a simple iterative approach. It is created the SVM model in linear CPU time of the dataset.

Advantages:

- High bandwidth and low latency
- Significantly improve the real-time experience and satisfaction of users.
- Edge computing nodes are scattered

IV. LITERATURE SURVEY

A) Title : A Machine Learning Perspective for Predicting Chronic Kidney Disease
Proposed By: Vanathi. D,S.M. Ramesh, Tamizharasu. K, Sengottaiyan. N
Published in : 20 August 2024
Summary : The proposed system uses machine learning algorithms like KNN, SVM, and ANN along with ensemble methods (Random Forest, Extra Trees, AdaBoost, XG Boost) for CKD prediction. It achieves 99.2% accuracy, improving early detection and management.
B) Title: Prediction of Chronic Kidney Disease Using Various Machine Learning Algorithms
Proposed By: Daravath Anil, Shaik Naimudden, Aujugari Santosh Reddy, A Lavanya
Published in : 20 April 2024
Summary: The proposed system predicts Chronic Kidney Disease (CKD) using machine learning by reprocessing clinical data handling missing values, and applying collaborative filtering.Multiple algorithms are used to improve accuracy and prevent overfitting.

V. METHODOLOGIES

The development of a predictive model for Chronic Kidney Disease (CKD) was approached through a structured, five-step methodology to ensure accuracy and reliability in classification. Each stage was designed to handle a specific aspect of the machine learning pipeline, from data preparation to performance evaluation.

1. Data Preprocessing

The dataset used in this study was sourced from the UCI Machine Learning Repository, which includes patient records with various medical attributes related to kidney function. Preprocessing involved handling missing values, normalizing numerical features, and converting categorical variables into a usable format. These steps ensured that the data was clean and consistent before being fed into machine learning models.

2. Feature Selection

To improve model efficiency and accuracy, several feature selection techniques were applied:

- **Correlation-Based Feature Selection:** Identifies features with strong relationships to the target variable.
- **Wrapper Method:** Evaluates subsets of features by training models and selecting the best-performing combinations.
- **LASSO Regression:** Least Absolute Shrinkage and Selection Operator was used to penalize less important variables, effectively reducing the number of input features while maintaining model performance.

3.Model Training

Seven machine learning classifiers were used for comparative analysis:

1) Artificial Neural Network (ANN)
2) C5.0 Decision Tree

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.12505

3) CHAID (Chi-square Automatic Interaction Detector)
4) Logistic Regression
5) Linear Support Vector Machine (LSVM) with L1 and L2 penalties
6) Random Tree
7) Each model was trained using different combinations of full and selected features to evaluate how input dimensionality affected accuracy and performance.

4. Handling Class Imbalance with SMOTE

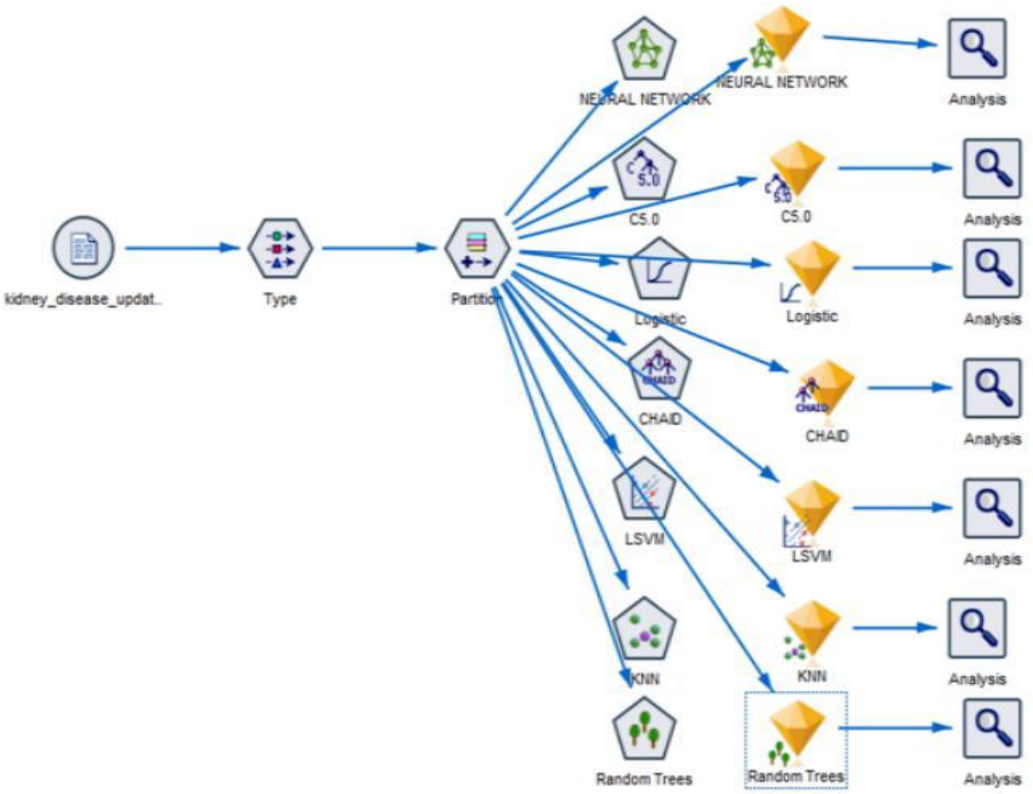
To address the class imbalance issue in the dataset (where CKD-positive and CKD-negative cases were not equally represented), the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This technique generates synthetic samples from the minority class to create a balanced dataset, improving the model's ability to correctly identify underrepresented cases.

5. Model Evaluation

The trained models were evaluated using key performance metrics including:

- Accuracy
- Precision
- Recall
- F1 Score
- Area Under the Curve (AUC)
- GINI Coefficient

VI. SYSTEM ARCHITECTURE



VII. CONCLUSION

The increasing prevalence of Chronic Kidney Disease (CKD) highlights the urgent need for accurate and timely diagnosis. This study demonstrates how machine learning, particularly Linear Support Vector Machines (LSVM) and Deep Neural Networks, can significantly enhance the detection process. By leveraging various feature selection

DOI: 10.17148/IARJSET.2025.12505

methods and addressing data imbalance with techniques like SMOTE, we achieved remarkably high prediction accuracy — with LSVM reaching up to 98.86% and the deep neural network pushing the boundary further to 99.6%.

These results show the practical potential of AI-driven tools in medical diagnostics, offering healthcare professionals a powerful support system for early intervention and improved patient outcomes. While challenges like computational complexity and edge computing limitations exist, the advantages in speed, accuracy, and real-time performance make machine learning a promising approach in the fight against CKD. Continued research and technological advancements will only strengthen these capabilities in the future.

These findings strongly suggest that machine learning, especially when combined with smart preprocessing and optimization strategies, can be a powerful tool in the early diagnosis of CKD. The study also illustrates the importance of choosing the right model and techniques based on the nature and quality of the data.

REFERENCES

[1] A. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017.
[2] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Pearson Education, 2009.
[3] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996.
[4] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 29, no. 2, pp. 119–127, 1980.
[5] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Wiley, 2013.
[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[7] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[10] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall, 1989.
[11] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, 1993.
[12] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997.
[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.
[14] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 245–271, 1997.
[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.