Data Analysis and Prediction of Chronic Kidney Disease Using Machine and Deep Learning Techniques

Dr. M. Rajendiran[1], Dr. M. Vijayakanth[2*]

[1]*Assistant Professor, Department of Computer Science, Government Arts and Science College, Jayankondam, Tamil Nadu, India.*
[2*]*Assistant Professor, Department of Computer Science, Thiru Kolanjiappar Govt.Arts College Virudhachalam, Tamil Nadu, India.*

Email: [1]rajendranmaha@gmail.com
Corresponding Email: [2]vijayakanth82@gmail.com*

KEYWORDS
Machine Learning, Chronic Kidney Disease, Machine Learning, Deep Learning, Performance Metrics. Decision Tree, Correlation Coefficient, and Test Statistics.

ABSTRACT:
The timely detection and effective management of chronic kidney disease (CKD) are essential for decelerating its progression and minimizing associated complications. Regular monitoring and strict adherence to medical guidance are vital for individuals diagnosed with CKD. Machine learning is a field dedicated to enabling computers to learn autonomously, thereby eliminating the need for explicit programming. Through this process, machines independently acquire knowledge by analyzing relevant data, reducing reliance on external inputs. As a cornerstone of deep learning, and neural networks, machine learning is pivotal for advanced pattern recognition and predictive modeling. This paper examines CKD-related datasets comprising attributes such as blood pressure (Bp), specific gravity (Sg), albumin (Al), sugar (Su), red blood cells (Rbc), blood urea (Bu), serum creatinine (Sc), sodium (Sod), potassium (Pot), hemoglobin (Hemo), white blood cell count (Wbcc), red blood cell count (Rbcc), hypertension (Htn), and disease classification (Class). Machine learning techniques, including Linear Regression, Multilayer Perceptron, SMOreg, M5P, Random Forest, REP Tree, and proposed deep learning approaches, are employed to analyze and predict this dataset. Numerical results, supported by statistical tests and accuracy metrics, are presented to validate the proposed methodologies.

1. Introduction and Literature Review

CKD, a widespread and severe health issue, impacts kidney function and frequently advances gradually. Investigative efforts into CKD encompass diverse dimensions, encompassing its origins, risk elements, diagnostic methods, treatment alternatives, management approaches, and broader influence on health. Data mining finds extensive utility across diverse industries such as marketing, finance, healthcare, and telecommunications. For instance, within marketing, it aids in delineating customer segments and tailoring marketing endeavors, whereas in healthcare, it assists in pinpointing disease risk factors and devising personalized treatment strategies. Machine learning finds application in a myriad of fields, spanning from image and speech recognition to natural language processing, recommendation systems, fraud detection, portfolio optimization, automated tasks, among others. Additionally, machine learning models play a pivotal role in enhancing the intelligence and adaptability of autonomous vehicles,

drones, and robots, enabling them to navigate and function effectively in dynamic environments.

The chronic kidney disease, and two others with features selected through Correlation Based Feature Selection and Chi-Square Test. Across all four cases, Sequential Minimal Optimization (SMO) and Multilayer Perceptron exhibited superior performance, achieving an average accuracy of 98.31% and 98.06%, respectively. These results, validated by F1-Score measurements exceeding 0.98 for both algorithms, underscore their efficacy. Classification rules were derived for all models, culminating in the extraction of action rules. The study underscores how comprehensive data analysis enables the creation of models that aid physicians in diagnosing diseases and guiding treatment decisions. These action rules serve as crucial directives for physicians, offering assurance in diagnoses and unveiling novel therapeutic pathways [1].

Chronic kidney disease (CKD) stands as a prevalent global affliction, claiming lives in significant numbers. Machine learning techniques for predicting chronic renal disease using clinical data, exploring two master teaching approaches: Random Forest Classifier (RFC) and Logistic Regression (LR). Utilizing the UCI dataset of chronic kidney disease components, we compared the results of these models to identify the most effective regression model for prediction. Among the two preprocessing scenarios—replacing missing values with column-wise mean values and selecting crucial features—the latter was deemed most logical, facilitating training with a more extensive dataset without data loss. However, correlation-based methods yielded the most favorable outcomes, achieving an accuracy of 98% in both scenarios. Consequently, this system can be implemented for early-stage CKD prediction in a cost-efficient manner, particularly beneficial for underdeveloped and developing countries [2].

The chronic kidney disease dataset, consisting of 25 distinct features, was obtained from the UCI Machine Learning repository. Analysis involved the utilization of three machine learning classifiers—Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM). To improve the model's performance, a bagging ensemble method was subsequently applied. The machine learning classifiers were trained based on clustering within the chronic kidney disease dataset. After categorizing and assessing the Kidney Disease Collection using both categories and non-linear features, the Decision Tree classifier showcased the most promising performance, achieving an accuracy of 95.92%. Upon implementing the bagging ensemble method, the model's accuracy surged to its peak at 97.23% [3].

The immense volume of healthcare transaction data poses a challenge for traditional processing methods due to its complexity and size. Data mining offers the methodology and technological means to convert this extensive data into actionable insights for informed decision-making. The healthcare industry is generally information-rich, making manual handling impractical. Utilizing these vast datasets is crucial in data mining to extract valuable information and establish connections among attributes. Managing kidney disease is a complex endeavor demanding substantial expertise and understanding. It stands as a silent threat in developed nations and represents a significant contributor to the disease burden in developing countries. Within the healthcare sector, data mining primarily focuses on disease prediction using datasets. In this context, data mining classification techniques such as Decision trees, Artificial Neural Networks (ANN), and Naive Bayes were examined using the kidney disease dataset [4].

Predict kidney disease by employing various machine learning algorithms: Support Vector Machine (SVM), Multilayer Perceptron (MLP), Decision Tree (C4.5), Bayesian Network (BN), and K-Nearest Neighbor (K-NN). The primary objective is to compare these algorithms and determine the most effective one(s) based on multiple criteria. The study utilizes the "Chronic Kidney Disease" database implemented on the WEKA platform. Upon analyzing the experimental results, it becomes evident that MLP and C4.5 showcase the most promising

outcomes. However, a comparison using the Receiver Operating Characteristic (ROC) curve indicates that C4.5 stands out as the most efficient algorithm [5].

Data mining proves to be a valuable tool for analyzing large existing databases and uncovering previously unknown information. In this paper, a weather dataset is used, where specific days are represented as rows, weather conditions as attributes, and a class indicating golf suitability. Six classification algorithms, including Linear Regression (LR), Multilayer Perceptron (MP), SMOreg, M5P, random forest (RF), and REP tree, are employed to measure accuracy [6].

Chronic Kidney Disease represents a significant global public health concern, characterized by kidney damage often associated with conditions like diabetes. Some patients experience both diabetes and chronic kidney disease concurrently. This paper aims to forecast mortality and the progression of diabetic chronic kidney disease to its final stage. Various data mining techniques were employed for this purpose, including the extraction of rules using rough set theory. Diverse features were utilized from patients with diabetic chronic kidney disease to predict disease progression. Methods such as J48, Naïve Bayes, Bayesian Network, SVM, SMO, Bagging, Random Forest, and Multilayer Perceptron were applied. To evaluate performance, comparisons were made based on Recall, Precision, and F-Measure metrics. The experimental results indicate that Random Forest outperforms other methods, demonstrating superior performance in predicting disease progression [7].

Study involved analyzing data from CKD patients and devising a system capable of predicting the risk of developing CKD. We utilized data from 455 patients, incorporating both an online dataset sourced from the UCI Machine Learning Repository and a real-time dataset collected from Khulna City Medical College. Python, a high-level interpreted programming language, was employed in developing our system. The data was trained using a 10-fold Cross-Validation technique, and both Random Forest and Artificial Neural Network (ANN) algorithms were applied. The Random Forest algorithm achieved an accuracy of 97.12%, while ANN achieved 94.5%. This system is poised to aid in the early detection of chronic kidney diseases, offering the potential for early intervention and management [8].

Chronic Kidney Disease (CKD) is a form of long-lasting ailment characterized by slow progression over time, persisting for an extended duration. At its advanced stage, it becomes life-threatening and requires either kidney replacement or regular dialysis, serving as an artificial filtering mechanism. Early identification of CKD is crucial to administer necessary treatments aiming to prevent or manage the disease effectively. This paper primarily concentrates on classification techniques, specifically examining tree-based decision trees, random forest, and logistic regression methods. Various measures have been employed to compare these algorithms using a dataset sourced from the standard UCI repository [9].

Efficient assessment of groundwater levels, rainfall, population, food grains, and enterprise data through stochastic modelling and data mining approaches. This approach is shown to effectively predict groundwater levels [10] and [11]. The paper analyzes chronic disease data, with specific location data represented as rows and attributes encompassing topics, questions, data values, low confidence limits, and high confidence limits. The study evaluates five classification algorithms, and the M5P decision tree approach is identified as the most effective for building models compared to other decision tree approaches [12].

2. Backgrounds and Methodologies

The tree's leaf nodes represent the final decision or the predicted outcome. The "CART" (Classification and Regression Trees) algorithm is the most used algorithm for building decision trees [13].

2.1 Linear Regression

Linear regression is a statistical technique employed to comprehend and forecast the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition.

Inputs

- X: Feature values (independent variable)
- Y: Target values (dependent variable)
- n: Number of data points

Steps

1. **Calculate the means** of X and Y:
   X_mean = (1/n) * sum(X_i) and Y_mean = (1/n) * sum(Y_i)
2. **Compute the slope (mmm)** of the line:
   m = sum((X_i - X_mean) * (Y_i - Y_mean)) / sum((X_i - X_mean)^2)
3. **Determine the y-intercept (b)**:
   b = Y_mean - m * X_mean
4. **Formulate the regression equation**:
   Y_hat = m * X + b
5. **Predict values** using the regression equation for given input X.

Outputs

Regression line: Y=mX+b

2.2 Multilayer Perception

A Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of interconnected nodes or neurons. It's a fundamental architecture in deep learning and is used for various tasks, including classification, regression, and more complex tasks like image recognition and natural language processing.

1. Initialize: Define the structure of the MLP:

- Input layer size (number of features).
- Number of hidden layers and neurons in each layer.
- Output layer size (number of classes or target variables).
- Initialize weights and biases randomly for all layers.
- Set hyperparameters: Learning rate, Number of epochs, and Batch size.

2. Forward Propagation: Pass input data through the network:

- Multiply inputs by weights, add biases, and apply an activation function at each layer.
- Continue this process layer by layer until the output layer is reached.

3. Compute Loss: Compare the predicted output with the actual target values using a loss function, such as:

- **Cross-entropy loss** for classification.
- **Mean squared error** for regression.

4. **Backpropagation:** Calculate the error at the output layer. Propagate this error backward through the network, adjusting the weights and biases of each layer based on the error contribution from that layer.
5. **Update Weights:** Update the weights and biases using an optimization method like gradient descent.
   - Adjust the weights to minimize the loss based on the calculated gradients.

6. **Repeat:** Steps 2–5 for the specified number of epochs or until the model converges (loss stops decreasing significantly).

7. **Predict:** After training, use forward propagation on new input data to make predictions.

2.3 SMO

Sequential Minimal Optimization (SMO) is an algorithm designed for training support vector machines (SVMs), a machine learning model widely applied in classification and regression tasks. It is particularly adept at addressing the quadratic programming problem encountered during SVM training. By dividing the optimization process into smaller, manageable subproblems, SMO focuses on two Lagrange multipliers simultaneously, streamlining the overall solution.

Input

- Dataset D = $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ where x_i are feature vectors and $y_i \in \{-1, 1\}$.
- Kernel function K (x_i, x_j).
- Regularization parameter C.
- Tolerance ϵ.
- Maximum number of iterations.

Output

- Optimized Lagrange multipliers $\alpha_1, \alpha_2, \ldots, \alpha_m$.
- Bias term b.

2.4 M5P

The M5P algorithm is a machine learning method for creating model trees, which combine decision trees and regression models. The algorithm constructs a tree structure where leaves contain linear regression models instead of constant values. Here's a simplified description:

M5P Algorithm

Input

- Training dataset $D=\{(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)\}$, where xi is a feature vector and y_i is the target value.

Output

- A model tree capable of predicting y for new x.

2.5 Random Forest

Random Forest is a popular machine learning ensemble method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy,

robustness, and ability to handle complex datasets. Random Forest is widely used in various domains, including data science, machine learning, and pattern recognition.

Bootstrap Sampling: For each tree t (from 1 to T):
- Create a bootstrap sample by randomly sampling n instances with replacement from the dataset.

Tree Construction: Train a decision tree on the bootstrap sample:
1. At each split in the tree:
   - Randomly select m features from the total features.
   - Find the best split point among the m features using a suitable criterion (e.g., Gini impurity for classification, variance reduction for regression).

2. Grow the tree to its maximum depth or until a stopping condition is met (e.g., minimum number of samples at a leaf node).

Prediction: For a new input x:
- For classification: Aggregate the predictions of all trees using majority voting.
- For regression: Compute the average of the predictions from all trees.

Output: Return the ensemble of decision trees.

2.6 REP Tree

The Reduced Error Pruning Tree (REP Tree) is an efficient decision tree algorithm designed for both classification and regression tasks. By employing the reduced error pruning method, it mitigates overfitting, resulting in a streamlined and more generalizable model suitable for diverse datasets.

Input
- Training dataset $D=\{(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)\}$, where x_i are feature vectors and y_i are target values.
- Validation dataset for pruning.
- Splitting criterion (e.g., Information Gain, Gini Index for classification; variance reduction for regression).

Output
- A pruned decision tree.

2.7 Proposed DL (FNNES): Feedforward Neural Network (FNN) with Early Stopping (ES)

To create a Deep Learning model for analyzing data and making predictions using Chronic Kidney Disease (CKD) data, a standard workflow for training and evaluating neural networks should be followed. Below is a detailed framework for constructing such a model. The objective is to predict the presence of Chronic Kidney Disease (CKD) in a patient, leveraging a range of medical variables. The CKD dataset typically comprises features like age, blood pressure, specific gravity, albumin, glucose levels, blood urea, serum creatinine, sodium, potassium, hemoglobin, and others. The model can be a feedforward neural network (FNN), which consists of several layers:

Model Architecture

1. **Input Layer**: The size of the input layer is determined by the number of features in the dataset (e.g., for the Chronic Kidney Disease (CKD) dataset, there are typically 24 features).
2. **Hidden Layers**: Usually, the model employs 2-3 hidden layers utilizing ReLU (Rectified Linear Unit) activation functions. While additional hidden layers can be explored, 2-3 are often adequate for this type of classification task.
3. **Output Layer**: For binary classification (CKD vs. non-CKD), the output layer should use a sigmoid activation function. For multi-class classification (such as predicting CKD stages), a softmax function is preferred.
4. **Dropout Layers**: To reduce the risk of overfitting, dropout layers should be incorporated after each hidden layer with a dropout rate ranging between 0.2 and 0.5.

Model Training

- **Epochs**: Define the number of epochs (complete passes through the dataset), commonly set to 50 or 100 for initial experimentation.
- **Batch Size**: Set an appropriate batch size, often chosen as 32 or 64.
- **Validation**: Utilize a validation set to monitor model performance during training, helping prevent overfitting by ensuring the model generalizes well to unseen data.

Model Evaluation

- **Accuracy**: After training, assess the model's accuracy on the test set to gauge its overall performance.
- **Confusion Matrix**: Construct a confusion matrix to visualize how effectively the model distinguishes between CKD and non-CKD instances.
- **Precision, Recall, and F1-Score**: These metrics are particularly significant in cases of imbalanced datasets, providing a deeper insight into the model's ability to correctly classify each category, beyond mere accuracy.

Model Tuning and Improvement

- **Hyperparameter Tuning**: Utilize methods such as GridSearchCV or RandomizedSearchCV to optimize hyperparameters, including learning rate, batch size, and the number of hidden units.

- **Advanced Techniques**
  - **Early Stopping**: Implement early stopping to terminate training when the validation accuracy ceases to improve, thereby preventing unnecessary computation and overfitting.
  - **Cross-Validation**: Employ cross-validation techniques to enhance model generalization and further reduce the likelihood of overfitting.

This framework outlines a comprehensive process for developing, training, evaluating, and fine-tuning a deep learning model for classification tasks, specifically designed to predict Chronic Kidney Disease or its stages.

2.7 Accuracy Metrics

The error rate of a predictive model can be assessed using various accuracy metrics commonly employed in machine learning and statistical analysis. In regression analysis, the core idea behind accuracy evaluation involves comparing the actual target values with the predicted ones [14]-[16].

Correlation Coefficient (r): The correlation coefficient (r) quantifies the degree of association between the dependent and independent variables. A value approaching 1 signifies a robust

positive relationship, whereas a value near -1 indicates a very strong negative correlation of association. A coefficient of 0, on the other hand, suggests the absence of any linear correlation between the variables.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] - [n\sum y^2 - (\sum y)^2]}}$$

Accuracy: The ratio of correct predictions to the total number of predictions made. Measures how often the model makes correct predictions. For deep learning, accuracy is commonly used in classification tasks, especially when the dataset is balanced. Machine learning is widely applied in traditional classification models such as decision trees, random forests, and SVM.
Accuracy = (TP + TN) / (TP+TN+FP+FN)
Where: TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives
Precision: The proportion of positive predictions that are actually correct. Useful when the cost of false positives is high (e.g., spam detection, disease diagnosis). Deep Learning: Applied in tasks like image classification and medical diagnosis. Machine Learning: Frequently used in models like logistic regression, decision trees, and random forests.
Precision = TP / (FP + TP)
Recall (Sensitivity or True Positive Rate): The proportion of actual positive cases that are correctly predicted by the model. Important when minimizing false negatives is critical (e.g., cancer detection, fraud detection). Deep Learning: Used in tasks where all positive instances should be identified (e.g., anomaly detection). Machine Learning: Applied in situations where missing a positive case is costly (e.g., fraud detection).
Recall = TP / (TP+FN)
F1-Score: The harmonic means of precision and recall, providing a balance between the two metrics. Especially valuable for imbalanced datasets, where precision and recall need to be optimized together. Deep Learning: Used when dealing with class-imbalanced data or in cases where both false positives and false negatives are costly.
F1=2 × ((Precision × Recall) / (Precision + Recall))

3. Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository. The chronic kidney disease dataset includes 14 parameters which have different categories of data like Bp, Sg, Al, Su, Rbc, Bu, Sc, Sod, Pot, Hemo, Wbcc, Rbcc, Htn, Class [17]. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. Chronic Kidney Disease Sample Dataset

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
80	1.02	0	0	1	22	1.2	139	4.6	16.5	4700	4.6	0	0
60	1.025	0	0	1	57	3.07	137	4.7	14	4500	5.5	0	0
60	1.025	0	0	1	46	1	135	5	15.7	6300	4.8	0	0
60	1.02	0	0	1	44	1.2	142	4.9	14.5	9400	6.4	0	0
70	1.025	0	0	1	23	0.6	140	4.7	16.3	5800	5.6	0	0
80	1.02	1	0	1	33	0.9	144	4.5	13.3	8100	5.2	0	0

Table 2: Machine Learning Models with Correlation Coefficient

ML Approaches	Correlation Coefficient
SMOreg	0.8195
Linear Regression	0.8209
REP Tree	0.9359
Multilayer Perceptron	0.9400
M5P	0.9602
Random Forest	0.9798
FNNES	0.9895

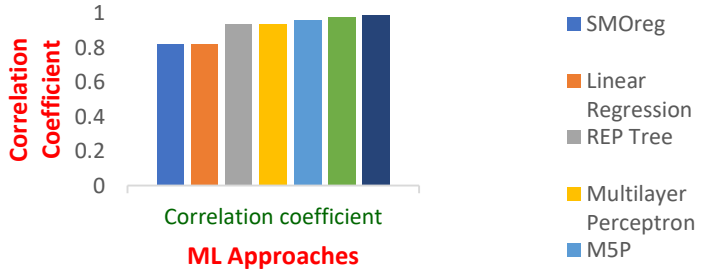

Figure 1. Correlation Coefficient of Parkinson's disease Data Analysis

Table 3. Performance Metrics for Parkinson's disease Data Analysis

Model/Algorithm	Accuracy	Precision	Recall	Specificity	F1-Score
SMO	84.45	82.89	85.12	83.56	83.24
Linear Regression	87.43	85.63	87.27	85.74	85.46
REP Tree	90.33	89.27	90.36	89.31	90.84
Multilayer Perceptron	92.67	90.48	92.85	90.21	91.63
M5P	93.86	91.11	93.96	91.10	92.22
Random Forest	94.12	92.24	94.78	92.40	93.74
FNNES	97.82	95.39	97.41	95.41	97.53

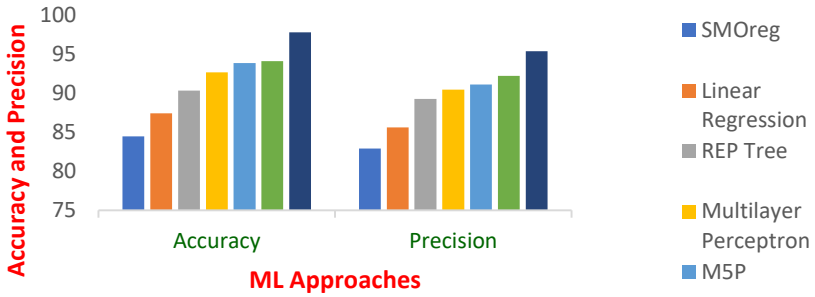

Figure 2. Accuracy and Precision of Parkinson's disease Data Analysis

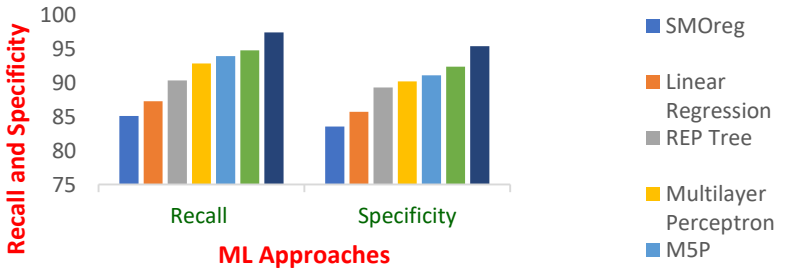

Figure 3. Recall and Specificity of Parkinson's disease Data Analysis

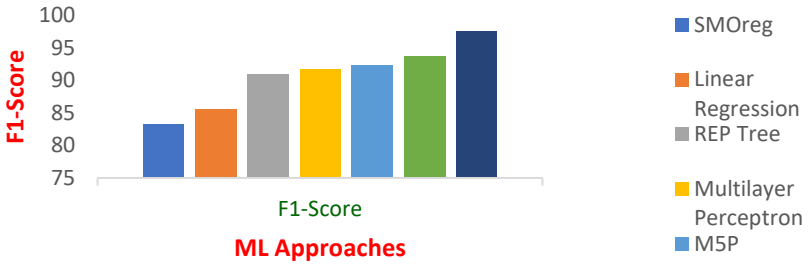

Figure 4. F1-Score of Parkinson's disease Data Analysis

4. Results and Discussion

The investigation scrutinizes the efficacy of diverse machine learning paradigms in forecasting Chronic Kidney Disease (CKD) by leveraging a dataset encompassing 14 intricate features, including parameters such as blood pressure, specific gravity, albumin, sugar levels, and red blood cell count. The empirical findings substantiate the effectiveness of the proposed methodologies, as elaborated below.

Table 1 delineates an exemplar of the CKD dataset, encompassing critical attributes such as blood pressure (Bp), specific gravity (Sg), albumin (Al), among others. These variables are pivotal in the nuanced process of CKD diagnosis and prediction, serving as foundational markers for analytical insights.

Model Performance: Table 2 encapsulates the correlation coefficients for the employed machine learning algorithms, highlighting that the FNNES model achieved the apex coefficient of 0.9895. This was closely trailed by the Random Forest model, recording 0.9798, and the M5P model at 0.9602. The coefficients are graphically depicted in Figure 1, accentuating the preeminence of FNNES in discerning patterns within the dataset with remarkable precision.

Comparative Analysis of Performance Metrics: A comprehensive comparison of performance metrics—spanning accuracy, precision, recall, specificity, and F1-score—is encapsulated in Table 3. Prominent observations include: The FNNES model, distinguished by its unparalleled accuracy of 97.82%, precision of 95.39%, and F1-score of 97.53%, which epitomizes its efficacy in CKD prediction. The Random Forest and M5P models demonstrated commendable performance, with accuracies of 94.12% and 93.86%, respectively. While the SMOreg model exhibited relatively modest metrics, its results remain indicative of reliable predictive capabilities.
Figures 2, 3, and 4 visually expound upon the accuracy, recall, specificity, and F1-scores of the models, reinforcing the dominance of deep learning approaches, particularly FNNES, over conventional machine learning techniques.

5. Conclusion

The results illuminate the transformative potential of sophisticated deep learning frameworks, particularly the FNNES model, in achieving unparalleled precision in CKD diagnosis, thereby eclipsing traditional machine learning counterparts. The substantial correlation coefficients and exceptional performance metrics underscore the utility of such models in enhancing clinical decision-making. Prospective endeavors could explore the integration of additional parameters or the development of hybrid models to further refine and augment predictive accuracy.

6. Reference

1. Turiac, A.S. and Zdrodowska, M., 2022. Data mining approach in diagnosis and treatment of chronic kidney disease. acta mechanica et automatica, 16(3), pp.180-188.
2. Babu, K.P. and Noorullah, S., 2022. Recognition of Chronic Kidney Disease Using Machine Learning. Journal of Algebraic Statistics, 13(1), pp.910-917.
3. Pal, S., 2022. Chronic Kidney Disease Prediction Using Machine Learning Techniques. Biomedical Materials & Devices, pp.1-7.
4. Bala, S. and Kumar, K., 2014. A literature review on kidney disease prediction using data mining classification technique.
5. Boukenze, B., Haqiq, A. and Mousannif, H., 2017. Predicting chronic kidney failure disease using data mining techniques. In Advances in Ubiquitous Networking 2: Proceedings of the UNet'16 2 (pp. 701-712). Springer Singapore.
6. Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the Weka tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.
7. Afhami, N., 2018. Prediction of diabetic chronic kidney disease progression using data mining techniques. International Journal of Computer Science Engineering, 7(2), pp.35-40.
8. Yashfi, S.Y., Islam, M.A., Sakib, N., Islam, T., Shahbaaz, M. and Pantho, S.S., 2020, July. Risk prediction of chronic kidney disease using machine learning algorithms. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.
9. Gupta, R., Koli, N., Mahor, N. and Tejashri, N., 2020, June. Performance analysis of machine learning classifier for predicting chronic kidney disease. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-4). IEEE.
10. Rajesh, P., Karthikeyan, M. and Arulpavai, R., 2019, December. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model. In AIP Conference Proceedings (Vol. 2177, No. 1). AIP Publishing.

11. Rajesh, P. and Karthikeyan, M., 2019. Data mining approaches to predict the factors that affect agriculture growth using stochastic models. International Journal of Computer Sciences and Engineering, 7(4), pp.18-23.
12. Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., 2019. Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), pp.1472-1477.
13. Kohavi, R., & Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).
14. Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/
15. S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, "Root mean square error (RMSE): A comprehensive review," International Journal of Applied Mathematics and Statistics, vol. 59, no. 1, pp. 42–49, 2019.
16. Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566
17. https://www.kaggle.com/datasets/abhia1999/chronic-kidney-disease/data