ISSN: 2633-4828 Vol. 6 No.1, January, 2024

*International Journal of Applied Engineering & Technology*

# STATE-OF-THE-ART MACHINE LEARNING FOR PREDICTION OF CHRONIC KIDNEY DISEASE

**Hamzeh Ghorbani*[1], Arsen Minasyan[2], Harutyun S. Hovhannisyan[3], Parvin Ghorbani[4], Simin Ghorbani[5], Mehrdad Babak Rad[6], Marieta Davidyan[7], Natali Minasian[8], Anahit Mkrtchyan[9], Eduard Avagyan[10], Mehdi Ahmadi Alvar[11] and Ali Ghazanfari[12]**

[1,2,7, 9,10] Faculty of General Medicine, University of Traditional Medicine of Armenia (UTMA), 38a Marshal Babajanyan St., Yerevan 0040, Armenia

[3]Department of Internal Disease Propaedeutics, Yerevan State Medical University, Yerevan, Armenia

[4]Department of Cardiology, Faculty of Medicine, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

[5]Department of Nursing and Midwifery, Faculty of Nursing, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran

[6]Department of Dentistry, University of Traditional Medicine of Armenia (UTMA), 38a Marshal Babajanyan St., Yerevan 0040, Armenia

[8]Faculty of General Medicine, Yaroslavl State Medical University, Yaroslavl, Russia

[11]Faculty of Engineering, Department of Computer Engineering, Shahid Chamran University, Ahwaz, Iran

[12]Department of General Medicine, Ulyanov Chuvash State University, Cheboksary, Russia

*Corresponding authors: Hamzeh Ghorbani (hamzehghorbani68@yahoo.com)

## ABSTRACT

*Chronic kidney disease (CKD) is a condition marked by a gradual decline in kidney function, with varying etiology. This study aims to predict CKD using 400 data points from an open dataset, focusing on key parameters related to diabetes and blood pressure, which are directly associated with CKD. The research utilizes various input variables, including diabetes mellitus, age, blood pressure, blood sugar levels, hypertension status, sodium levels, hemoglobin levels, albumin levels, red blood cell count, potassium levels, and creatinine levels.*

*To achieve this prediction, three powerful classification algorithms—Stochastic Gradient Descent (SGD), Functional Tree (FT), and Logistic Model Tree (LMT)—are used. The outcomes reveal the superior accuracy of SGD compared to the other algorithms, attributed to its efficiency with large datasets, rapid convergence, adaptability, and suitability for diverse classifications and batch processing. The study establishes a hierarchy of algorithm performance, with SGD ranking highest, followed by LMT and FT. In conclusion, the findings underscore the potential of the SGD algorithm for precise predictions in the context of CDK.*

***Keywords:*** *chronic kidney disease (CKD), prediction, classification algorithms, stochastic gradient descent algorithm (SGDa), data analysis.*

## INTRODUCTION

The kidney is one of the vital organs responsible for filtering toxins and regulating bodily osmosis, plays an important role in maintaining our health [1]. Chronic Kidney Disease (CKD) is a dangerous condition characterized by the decrease in kidney function [2]. This disease develops over several years, resulting in the kidney's inability to efficiently the blood filtration, leading to the accumulation of water and toxins within the body [3]. Early detection of CKD is important, as it can help prevent complications such as high blood pressure, nerve damage, heart disease, diabetes, reduced immune function, and imbalances in electrolytes [4].

The CKD presents a range of symptoms, including nausea, vomiting, abdominal pain, diarrhea, fever, skin rashes, and back pain, which may vary based on the patient's age and gender. Notably, CKD is closely associated with

two primary risk factors: diabetes and high blood pressure. Preventing these underlying conditions is paramount in averting the onset of CKD [5].

Several diagnostic tests are available to assess CKD, including:

- Urine test
- Estimation of glomerular filtration rate (GFR)
- Creatinine test (Cr)

By recognizing the early signs of CKD and addressing its risk factors, individuals can take proactive steps to maintain their kidney health and overall well-being.

The CKD progresses through five distinct categories, each of them are based on the GFR value. These categories serve as important condition in assessing the severity of CKD [6]:

- **Stag 1:** GFR <= 90 mL/min: In this initial stage, individuals may have an increased risk of developing CKD. Their GFR remains relatively high, indicating that kidney function is within normal limits.
- **Stag 2:** GFR 60-89 mL/min: At this stage, there is a mild reduction in GFR, suggesting a slight decrease in kidney function. Patients may still have few noticeable symptoms.
- **Stag 3:** GFR 30-59 mL/min: As CKD progresses, the GFR continues to decrease. This moderate reduction in GFR may lead to more pronounced symptoms, and it becomes essential to closely monitor kidney health.
- **Stage 4:** GFR 15-29 mL/min: At this stage, individuals experience a significant decrease in GFR, and kidney function is notably impaired. Patients may exhibit severe symptoms and complications, necessitating close medical management.
- **Stage 5:** GFR < 15 mL/min or dialysis: The final category represents CKD failure, where kidney function is severely compromised. At this point, patients often require dialysis or kidney transplant to manage the critical reduction in GFR and maintain overall health.

Understanding these stages and the associated GFR values is important for both healthcare professionals and individuals at risk of or living with CKD. Early detection and proper management can significantly improve the quality of life for those affected by this disease.

## MATERIALS AND METHODS

### Workflow

Workflow show fast review for whole of the article in few quickly. As shown in the workflow diagram, this approach consists of a structured set of steps that facilitate the process. This process includes:

**Data Collection:** The first step involves gathering data from open data, which is foundation for the analysis.

**Preprocessing:** This step focuses on removing outlier data high noise levels within the dataset. This important phase ensures that the data is clean and suitable for further analysis.

**Data Division:** The dataset is then divided into three distinct sections:

- **Training Data:** This data portion is for constructing and training the algorithms.
- **Testing Data:** This data portion of the dataset is for testing the algorithms.
- **Validation Data:** The validation dataset used for evaluating and accuracy of the developed algorithms.

Algorithm's evaluation and find best algorithm: The final step is the developed algorithms to the test data. This

phase assesses their performance and allows for the identification of the best-performing algorithm.

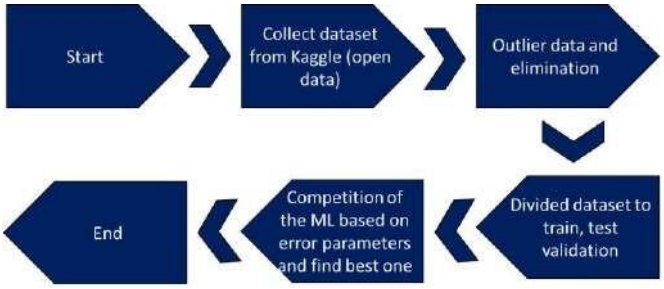


**Figure 1:** Workflow diagram for prediction of CKD.

Based on WHO's criteria, many factors can affect CKD. These parameters play an important role in the prediction of CKD. Table 1, sourced from an opensource data availibility, presents a comprehensive list of such influencing factors, with data collected from 400 data in the open data as part of this research worked.

To prediction of CKD, a range of parameters affecting the disease have been collected. These parameters include sex, age, blood pressure, blood sugar levels, blood urea nitrogen, hypertension status, sodium levels, hemoglobin levels, albumin levels, red blood cell count, potassium levels, and creatinine levels.

**Table 1:** Data variable for prediction of CDK.

| Abbreviation | Feature |
|---|---|
| DM | Diabetes mellitus |
| AG | Age |
| BP | Blood pressure |
| BSL | Blood sugar levels |
| BP | Blood pressure |
| HPT | Hypertension status |
| Na | Sodium levels |
| HGB | Hemoglobin levels |
| ALB | Albumin levels |
| RBC | Red blood cell count |
| K | Potassium levels |
| Cr | Creatinine levels |

**Data Preprocessing**

The important process involves the identification and handling of noisy data in the dataset and ensuring its reliability. During this process, we use outlier detection methods to eliminate outlier data points, enhancing the overall quality and suitability of the dataset for algorithm development and analysis. By filtering outlier data, we ensure the data can use for algorithm development.

**Stochastic Gradient Descent Algorithm (SGD)**

The SGD is a powerful optimization technique that simplifies the computation by utilizing a randomly selected, small subset of the training data [7]. The term "batch size" refers to the number of training data sets used to approximate the gradient in a single iteration. The SGD stands out by allowing parameter updates to be performed using this smaller batch size instead of the SGD method, resulting in faster convergence [8].

In the most extreme case, when the batch size is set 0 to 1, SGD delivers the maximum update frequency, resembling a simple perceptron-like algorithm.

In the context of SGD, the feature weights for each training sample are updated using the following Equation 1:

$$S^{P+1} = S^{P} + \beta_P \frac{\partial}{\partial S}\left(F(i,S) - \frac{D}{E}\sum_{i}|S_i|\right) \quad (1)$$

The SGD's efficiency makes it a valuable tool for optimization in various ML models [9]. By using smaller batch sizes and frequent updates, SGD accelerates the convergence of algorithms, enabling rapid adaptation to changing data. Figure 2, illustration graph for SGD.

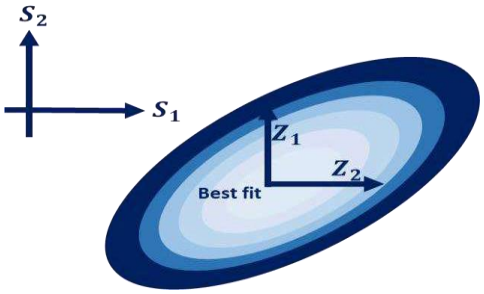


**Figure 2:** Illustration graph for SGD.

**Logistic Model Tree Algorithm (LMT)**

The LMT stands as a hybrid tree classifier that combines elements of decision trees and logistic regression techniques. Within the framework of LMT, the classification and regression tree algorithms are used for tree pruning in the context of classification function [10]. Additionally, LMT leverages the LogitBoost algorithm to construct logistic regression models at each tree node, enhancing its predictive [11].

The division process in LMT is guided by logistic-type information gain, a strategy known for its effectiveness in optimizing tree structures. To determine the optimal number of iterations for the LogitBoost algorithm and decrease the risk of overfitting, LMT wisely use cross-validation [12]. Within the LMT framework, an incremental logistic regression approach, reminiscent of the least squares fit, is utilized within each class (Mi) (Equation 2) [13].

$$L_M(x) = \sum_{i}^{n} \gamma_i x_i + \beta_0 \quad (2)$$

LMT's integration of decision tree and logistic regression techniques, combined with strategic model construction, renders it a versatile tool for a wide array of classification tasks. It capitalizes on the strengths of both methods to create an effective and accurate predictive model. Figure 3, illustration graph for LMT.

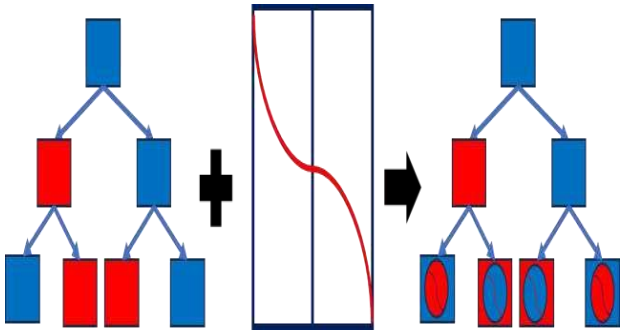

**Figure 3:** Illustration graph for LMT.

**Functional Tree Algorithm (FT)**

A FT serves as a tree classifier that harnesses a combination of features found within the leaf nodes, decision nodes, or both, as well as within the leaves of the learning classification tree [14]. In the FT framework, logistic regression functions are used to partition the tree into functional internal nodes, enabling the prediction of functional leaves [14]. Functional leaves within the FT play an important role in variance reduction, while the functional internal nodes are instrumental in minimizing classification bias [15]. Notably, the application of FT in the context of landslide prediction has been somewhat limited, with only a few case studies to date [16]. The algorithm for FT classification is a series of key steps: (1) constructing the model, which involves determining the probability distribution of output classes through the selection of a linear Bayes discriminant function constructor; (2) generating a new dataset by expanding the new factors associated with landslide or non-landslide classes; and (3) constructing the classification tree by choosing factors from both the original training dataset and the new dataset [17]. Figure 4, illustration graph for FT.

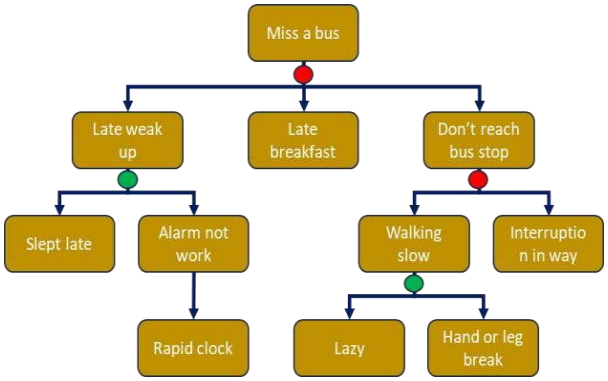


**Figure 3:** Illustration graph for FT.

**DISCUSSION OF RESULTS**

To evaluate and compare the performance of classification algorithms for predicting CKD, statistical parameters play an important role. Equations 3-6 define these essential statistical metrics. These metrics serve as valuable tools for determine the accuracy and effectiveness of different algorithms, ultimately helped the selection of the most suitable approach.

These statistical parameters encompass a range of important measures, such as precision (Pr), recall (Re), F1 score, and accuracy (AC), which enable a comprehensive analysis of algorithm performance. By quantifying the algorithms' ability to correctly identify CKD cases and non-CKD cases, these metrics empower researchers and healthcare professionals to make informed decisions regarding the choice of the most effective algorithm for CKD prediction.

$$Ac = \frac{TP_{CKD} + TN_{CKD}}{TP_{CKD} + FP_{CKD} + TN_{CKD} + FN_{CKD}} \times 100 \quad (3)$$

$$Re = \frac{TP_{CKD}}{TP_{CKD} + FN_{CKD}} \times 100 \quad (4)$$

$$Pr = \frac{TP_{CKD}}{TP_{CKD} + FP_{CKD}} \times 100 \quad (5)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

Table 2 provides a result of the algorithms used to predict CKD, illustrating the results obtained for the test, training, and validation sets. An analysis of the outcomes shows differences in performance accuracy among these algorithms.

The data shows the SGD algorithm outperforms the other two algorithms, LMT and FT, in terms of performance accuracy. This result shows the SGD algorithm has higher accuracy than other algorithms which used for prediction of CDK.

**Table 2:** The ML results based on statical metric for prediction of CDK.

| Dataset | Models | F1 | Ac | Pr | Re |
|---|---|---|---|---|---|
| Trian | SGD | 0.97 | 0.95 | 0.96 | 0.96 |
| | LMT | 0.84 | 0.85 | 0.84 | 0.84 |
| | FT | 0.73 | 0.72 | 0.73 | 0.73 |
| Test | SGD | 0.94 | 0.93 | 0.94 | 0.94 |
| | LMT | 0.82 | 0.83 | 0.82 | 0.82 |
| | FT | 0.7 | 0.69 | 0.69 | 0.69 |
| Validation | SGD | 0.93 | 0.92 | 0.93 | 0.93 |
| | LMT | 0.83 | 0.84 | 0.83 | 0.83 |
| | FT | 0.72 | 0.71 | 0.72 | 0.72 |

Two important statistical parameters for comparing classification algorithms are Recall (Re) and Precision (Pr). These metrics show the accuracy of the algorithms. By comparing these parameters, we can evaluate the algorithm's performance.

Based on the results shown in Table 2 and Figure 5, it is clear that the accuracy of the algorithms applied in this dataset is SGD > LMT > FT. These results emphasize the relative merits of these algorithms in correctly identifying and classifying samples, with SGD showing the highest accuracy, followed by LMT and FT, respectively. Understanding and using statistical parameters such as Recall and Precision in the evaluation process is important for making informed decisions when choosing classification algorithms. These metrics allow for a deeper understanding of the algorithm's ability to correctly classify samples and optimize its performance in real-world applications.

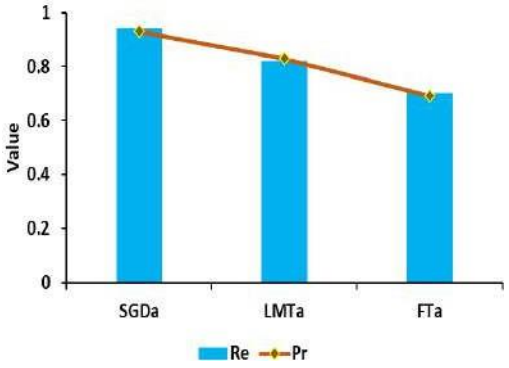


**Figure 5:** Combine chart based on Re and Pr statical classification for prediction of CDK.

When evaluating and comparing classification algorithms, two of the most important statistical parameters are AC and the F1 score. These metrics play an important role in assessing an algorithm's effectiveness, offering insights into its performance and predictive capabilities.

By definition, these algorithms' accuracy can be readily ascertained through these parameters, ensuring a comprehensive evaluation of their performance.

In addition, an analysis of the data presented in Table 2 and the insights gleaned from Figure 6 clearly indicates that SGD outperforms both LMT and FT in terms of performance accuracy in predicting CDK. This observation emphasizes SGD's superior ability to correctly classify CDK cases, making it a promising choice for CDK prediction compared to the other algorithms.

Based on these statistical parameters, such as Accuracy and the F1 score, is essential when assessing and selecting classification algorithms.

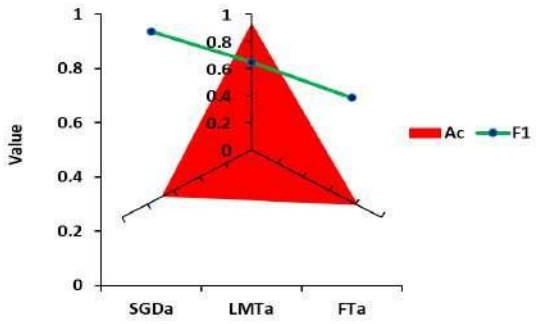


**Figure 6:** Combine chart based on Ac and F1 statical classification for prediction of CDK.

## CONCLUSION

The kidney is avital organ which is responsible for detoxification and osmotic regulation for the heath. Chronic Kidney Disease (CKD) marks a formidable health challenge, characterized by the insidious decline of renal function, progressing steadily over the course of several years. In this study, the information of 400 data from open data is used, the purpose of which is to predict CDK. This parameter is a very important parameter that is directly related to diabetes and blood pressure. In this research, various variables have been used for input, including: diabetes mellitus (DM), age (AG), blood pressure (BP), blood sugar levels (BSL), blood pressure (BP), hypertension status (HPT), sodium levels (Na), hemoglobin levels (HGB), albumin levels (ALB), red blood cell

count (RBC), potassium levels (K), and creatinine levels (Cr). In order to predict CDK, three powerful classification algorithms Stochastic Gradient Descent Algorithm (SGD), the Functional Tree Algorithm (FT), and the Logistic Model Tree Algorithm (LMT) have been used. The results of this research show the high accuracy of SGD algorithm compared to other algorithms. Among the advantages of the SGD algorithm are: productivity with large data sets, fast convergence, compatibility, and suitability for different classifications, batch processing. This research shows that SGD algorithm has higher performance accuracy than LMT and FT. Consequently, this study ranks the algorithms based on their performance accuracy as follows: SGD > LMT > FT.

## DATA AVAILIBILITY

The research was conducted with permission granted by the at the open data source of the Kaggle (Kaggle Dataset: https://www.kaggle.com/datasets/colearninglounge/chronic-kidney-disease?resource=download).

## CONFLICT OF INTEREST

The authors don't have any conflict of interest with this article.

## REFERENCES

[1] J. V. Bonventre, V. S. Vaidya, R. Schmouder, P. Feig, and F. Dieterle, "Next-generation biomarkers for detecting kidney toxicity," *Nature biotechnology*, vol. 28, no. 5, pp. 436-440, 2010.

[2] B. Gudeti, S. Mishra, S. Malik, T. F. Fernandez, A. K. Tyagi, and S. Kumari, "A novel approach to predict chronic kidney disease using machine learning algorithms," 2020, pp. 1630-1635: IEEE.

[3] M. Faria and M. N. de Pinho, "Challenges of reducing protein-bound uremic toxin levels in chronic kidney disease and end stage renal disease," *Translational Research*, vol. 229, pp. 115-134, 2021.

[4] M. Ala, "SGLT2 inhibition for cardiovascular diseases, chronic kidney disease, and NAFLD," *Endocrinology*, vol. 162, no. 12, p. bqab157, 2021.

[5] V. A. Luyckx *et al.*, "Reducing major risk factors for chronic kidney disease," *Kidney international supplements*, vol. 7, no. 2, pp. 71-87, 2017.

[6] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *The lancet*, vol. 389, no. 10075, pp. 1238-1252, 2017.

[7] O. Larsson, "Robustness, Stability and Performance of Optimization Algorithms for GAN Training," 2021.

[8] L. Luo, H. Ye, Z. Huang, and T. Zhang, "Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20566-20577, 2020.

[9] S. Baharlouei and M. Razaviyayn, "Dr. FERMI: A Stochastic Distributionally Robust Fair Empirical Risk Minimization Framework," *arXiv preprint arXiv:2309.11682*, 2023.

[10] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine learning*, vol. 59, pp. 161-205, 2005.

[11] D. Dancey, Z. A. Bandar, and D. McLean, "Logistic model tree extraction from artificial neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 4, pp. 794-802, 2007.

[12] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," 2005, pp. 675-683: Springer.

[13] S. A. Fayaz, M. Zaman, and M. A. Butt, "An application of logistic model tree (LMT) algorithm to ameliorate Prediction accuracy of meteorological data," *International Journal of Advanced Technology and*

*Engineering Exploration,* vol. 8, no. 84, pp. 1424-1440, 2021.

[14] J. Gama, "Functional trees," *Machine learning,* vol. 55, pp. 219-250, 2004.

[15] J. Gama, "Functional trees for classification," 2001, pp. 147-154: IEEE.

[16] L. Canete-Sifuentes, R. Monroy, and M. A. Medina-Perez, "FT4cip: A new functional tree for classification in class imbalance problems," *Knowledge-Based Systems,* vol. 252, p. 109294, 2022.

[17] A. Joshuva and V. Sugumaran, "A data driven approach for condition monitoring of wind turbine blade using vibration signals through best-first tree algorithm and functional trees algorithm: A comparative study," *ISA transactions,* vol. 67, pp. 160-172, 2017.