Information & Security
Published since 1998

W. Almayyan & B. AlGhannam
vol. 55, no. 2 (2024): 199-212
https://doi.org/10.11610/isij.5508
ISSN 0861-5160 (print), ISSN 1314-2119 (online)

Research Article

A Hybrid Bat-Grey Wolf Algorithm for Predicting Chronic Kidney Disease

Waheeda Almayyan* (✉) *and Bareeq AlGhannam

The Public Authority for Applied Education and Training, College of Business Studies, Kuwait, https://e.paaet.edu.kw/EN/Pages/default.aspx

ABSTRACT:

Chronic Kidney Disease (CKD) is a significant global health concern characterized by major global health issues defined by progressive kidney damage. Early detection is essential for chronic kidney disease (CKD) as it often progresses silently and can lead to severe complications. This study introduces a novel hybrid optimization algorithm, combining the Bat Algorithm (BA) and Grey Wolf Optimizer (GWO), to address the challenges of CKD prediction. By leveraging BA's exploration capabilities and GWO's exploitation potential, the proposed algorithm effectively searches for optimal solutions, enhancing the accuracy and efficiency of CKD diagnosis. The proposed algorithm, termed BAGWO, iteratively refines solutions through a two-stage process, enhancing the model's ability to converge on accurate and informative results. This algorithm not only enhanced classification accuracy but also reduced the feature set to only six attributes. The Random Forest classifier resulted in a recognition rate of 99.5 %, making it the top-performing classifier.

ARTICLE INFO:

RECEIVED: 26 AUG 2024
REVISED: 21 OCT 2024
ONLINE: 31 OCT 2024

KEYWORDS:

chronic kidney disease, CKD prediction, bat algorithm, grey wolf optimizer, hybrid optimization, feature selection, machine learning.

Introduction

The progressive nature of CKD is attributed to the kidneys' incapacity to remove waste products from circulation. CKD is a silent yet critical health threat affecting millions worldwide.[1] This progressive condition can lead to severe complications, including heart disease, stroke, and increased mortality. Early detection

✉ E-mail: wi.almayyan@paaet.edu.kw

is crucial as CKD often progresses without noticeable symptoms.[2] The prevalence of CKD as a significant global health issue necessitates advanced diagnostic tools. Early detection is crucial due to the often asymptomatic nature of CKD and its potential for severe complications.[3]

The application of machine learning (ML) has become increasingly effective in tackling complex healthcare problems, including CKD.[4] By leveraging advanced data mining techniques and sophisticated algorithms, researchers aim to improve early detection and patient outcomes.

Regarding the prediction of CKD using patient medical information, there is a sizable gap in the studies. The present study uses data mining techniques to identify predicted patterns in complex healthcare data in an attempt to overcome this restriction. To overcome the challenges posed by high-dimensional data, feature selection methods, as advocated by Wang and Alexander [5] are employed to identify relevant information. The goal is to develop a robust predictive model for early CKD diagnosis using the University of California Irvine (UCI) Machine Learning Repository dataset.

The goal of this research is to create a reliable predictive model that will aid in the early identification of CKD and enhance patient outcomes. Our goal is to find significant indicators of chronic kidney disease (CKD) by using data mining techniques and feature selection. Our proposed model offers a promising approach to early disease detection, contributing to improved patient outcomes and healthcare. In this work, we provide a novel hybrid optimisation algorithm that combines the advantages of the GWO and the Bat BA. While BA excels at local exploration, GWO's balanced exploration and exploitation capabilities make it well-suited for global optimization. The hybrid approach leverages BA's initial search for promising solutions, followed by GWO's targeted refinement. This synergistic combination enhances the algorithm's capability to identify optimal solutions.

The rest of the article is organised as follows: A detailed review of the literature on CKD prediction methods is provided in Section 2. The suggested technique, including feature selection, pre-processing of the data, and classification methods, is described in Section 3. The experimental findings are shown in Section 4, along with a comparison of the suggested approach with other state-of-the-art techniques. Section 5 concludes by discussing the limits of the study and the implications of the results.

Review of Related Literature

Several studies have successfully applied ML techniques to the UCI CKD dataset, leading to valuable insights into disease classification. By analyzing the diverse range of data within this dataset, researchers have developed predictive models that offer promise for early detection and diagnosis of CKD.

Swain et al.[6] developed a model employing feature scaling, balancing, and missing-value imputation techniques. A chi-square test informed the selection of nine features for classification using Support Vector Machines. Although the study added to the field, it was constrained by the lack of advanced methods for the imputation of missing values, which could have prevented the extraction of significant information.

Ebiaredoh-Mienye and Swart[7] developed an ML method integrating AdaBoost classifier with information-gain-based feature selection to predict the progression of CKD. In the study, this model's efficacy was contrasted with log regression, decision trees, Random Forest (RF), support vector machines, and traditional AdaBoost. Although the study showed the benefits of feature selection, it was constrained by the lack of data scaling and hyperparameter optimisation strategies, which may have improved model performance and generalisability.

Farjana et al.[8] employed multiple ML algorithms for CKD prediction, utilizing holdout validation and basic imputation techniques. While LightGBM demonstrated promising results, the absence of advanced data pre-processing and feature selection hinder the generalizability and robustness of the findings.

Islam et al.[9] employed ML techniques for CKD prediction, incorporating principal component analysis and recursive feature elimination for dimensionality reduction. While the study utilized basic imputation methods, it lacked advanced pre-processing techniques such as data scaling and hyperparameter optimization.

Hassan et al.[10] focused on predicting CKD using ML techniques. The study employed K-means clustering for data pre-processing and XGBoost and SHAP value analysis for feature selection. While the study incorporated advanced techniques for data pre-processing and feature selection, the absence of data scaling and hyperparameter optimization may limit the model's overall performance and generalizability.

Kaur et al.[11] employed several ML techniques, hybridizing Little's Missing Completely At Random (MCAR) to impute missing data and Ant Colony Optimization for feature selection. While the study employed ensemble methods, including bagging, it lacked critical pre-processing steps such as data scaling and hyperparameter optimization.

A deep neural network-based Multi-Layer Perceptron (MLP) classifier with a claimed flawless 100 % accuracy rate was proposed by Sawhney et al.[12] Although this is an outstanding result, the lack of crucial data pre-processing processes, like data scaling and missing value handling, compromises the study's methodological rigour. The robustness and generalisability of the model are seriously questioned in light of these omissions.

While promising, the present state of research on CKD prediction indicates a number of limitations that impede further advancement. This work aims to address these constraints directly by creating innovative methods. The main primary contribution of this work is the possibility to enhance healthcare by facilitating the early identification of individual instances of CKD. Before creating the ML model, this is achieved by selecting the features from the dataset that are most relevant and eliminating those that have lower contribution to prediction accuracy. By locating a compact subset of significant features from a large dataset, feature selection maximises model performance. Algorithms that draw inspiration from nature efficiently handle this problem by eliminating redundant and useless data.[13] By streamlining the model training process, this method lowers time complexity and computational overhead while maintaining accuracy.

Methods

Artificial intelligence has been applied to many challenging problems in the past few years, including clinical practice and public health.[13] Metaheuristic algorithms inspired by nature have proven to be successful in a range of diagnostic applications. These algorithms, inspired by biological phenomena, have been successfully employed to develop CKD diagnostic techniques.[14] By selecting a subset of key features, these algorithms enhance classification accuracy while reducing computational complexity. This approach addresses the challenge of high-dimensional data often encountered in ML tasks. Because of this, the training procedure has a sharp structure without sacrificing the projected accuracy attained by using only the most crucial characteristics. Thus, by using nature-inspired methods, the problem's computational cost and temporal complexity may be reduced.

This study compares various ML algorithms for predicting kidney disease. An accuracy rate of 90 % or higher was considered excellent.[4] By employing a diverse range of algorithms, this research surpasses previous studies, achieving a remarkable 99.5 % accuracy rate.

For the implementation of our intended model implementation, we relied on Weka API in Java. WEKA, a Java-based tool, provides a comprehensive suite of ML algorithms, including supervised and unsupervised techniques. It has a visualization environment, and it also supports the implementation of new ML algorithms.

CKD Dataset

The CKD medical dataset, comprised of 400 instances (250 CKD patients and 150 healthy controls), was collected at Apollo Hospitals over two months and subsequently donated to the UCI Machine Learning Repository on July 2, 2015, by Rubini, Soundarapandian, and Eswaran. Each instance includes 24 features, where 11 are numeric and 13 nominals, obtained from blood and urine samples, as well as patient surveys. More details and the specification of the dataset are provided by Khamparia et al.[15]

Data pre-processing

Data quality is paramount for effective data mining in the context of CKD prediction. To ensure accurate and reliable results, addressing missing values and inconsistencies within the dataset is essential. Data pre-processing techniques, including data integration, cleansing, reduction, transformation, and discretization, are crucial steps in transforming raw data into meaningful information. By meticulously preparing the data, we can optimize the performance of ML models and enhance the accuracy of CKD predictions. In this study, the CKD dataset underwent data cleaning to fill in all missing values, eliminate noise, and fix errors.

1. **Imputing the Missing Values**: A notable proportion of missing values was found in the CKD dataset. To address this, we employed imputation techniques. For nominal attributes, the mode value was used to fill in the missing values. For continuous attributes, the mean value was substituted. This approach preserved valuable data while maintaining data integrity.

2. **Data Splitting**: Ten-fold cross-validation was used to increase the efficacy of the model training process. Ten subsets of the CKD dataset were created; in each iteration, one subset was used for testing and the other nine for training. To prevent overfitting, this method makes use of a sizable amount of the data for training.

3. **Data Scaling**: To ensure consistent feature representation and prevent bias from features with varying scales, data scaling techniques were applied. This process transformed the data to a common scale, facilitating accurate comparisons and improving model performance.

Bat Optimization

The BA is a new metaheuristic algorithm addition that Yang has implemented.[16] The algorithm employs an echolocation-inspired search strategy, similar to how bats navigate using sound waves. To put it simply, bats utilise echolocation to navigate their environment. Bats can travel, find prey, and identify numerous objects around them, even in complete darkness. They achieve this by placing calls to the neighbourhood and keeping an ear out for any echoes that come back to them. They can identify other objects and determine their distance from them by listening to the delay in the return sound. Yang idealised bat echolocation behaviour to develop a numerical method for resolving optimisation problems.

In the Bat algorithm, virtual bats search for optimal solutions by randomly moving within the search space. Each bat is characterized by its position (X_i), velocity (V_i), frequency (F), and loudness (A_0).

In the Bat algorithm, each bat of the population represents a candidate solution. The bat with the best fitness value at the end of the optimization process is considered the optimal solution. Each candidate solution is illustrated with the help of vector $X_i=(X_1 \ldots X_i)^t$ with real value elements X_{ij},for i= 1…N_p and the interval for each element is taken from X_{ij} within the range [$X_{lb\ldots}$ X_{ub}]. While X_{lb} and X_{ub} determine the lower and upper bounds, N_p represents the size of the population. The bats adjust their positions based on these parameters, mimicking the echolocation behaviour of real bats.[17]

$$Fi = Fmin + (Fmax - Fmin)\beta \tag{1}$$

$$V_i{}^{t} = V_i{}^{t-1} + (X_i{}^{t} - X_*)F_i \tag{2}$$

$$X_i{}^{t} = X_i{}^{t-1} + V_i{}^{t} \tag{3}$$

where β is a random vector in the range [0, 1] drawn from a uniform distribution. X_* denotes the current global best solution. While the minimum and maximum frequency values are denoted by F_{min} and F_{max}, respectively. V_i is the velocity vector's symbol. A probabilistic local search is conducted using a random walk, as defined by the equation:

$$X_{new} = X_* + \varepsilon A_i^t \tag{4}$$

where A_i^t represents the average loudness of the bat population at time t, and ε is a random number uniformly distributed between -1 and 1. The loudness is updated iteratively using the equation:

$$A_i^{t+1} = \delta A_i^t \tag{5}$$

where δ is an experimentally determined constant.
Where r_i^0 is the initial pulse emission rate and is a constant greater than 0. The local search is controlled by the emission rate r_i, which is updated by the following equation

$$r_i^{t+1} = r_i^0[1 - \exp(-\gamma^t)] \tag{6}$$

Grey Wolf Optimization Algorithm

The natural leadership structure and hunting tactics of grey wolves served as the inspiration for researcher Mirjalili's 2014 introduction of the GWO idea.[18] Encircling the prey, assaulting the prey, and maintaining the social order are the three key components of the optimisation process. As an example, the grey wolf optimisation problem has no particular constraints on the objective function and is not dependent on rigorous mathematical features. Moreover, putting these procedures into practice is not difficult. Given that it can create a workable algorithm implementation plan in response to specific situations, the GWO is remarkably flexible.[19]

The nearby wolves (solutions) to the optimal solution are designated as alpha (α), beta (β), and delta (δ) wolves, respectively, representing the top three solutions within the population.

Their positions in the search space are represented by X_α, X_β, and X_δ. The remaining wolves follow these leaders, updating their positions based on their locations:

$$X(t + 1) = X_p(t) - A.|C.X_p(t) - A.X(t)| \tag{7}$$

where t represents the current iteration, A and C are coefficient vectors, X_p is the position vector of the prey, and X denotes the position vector of a grey wolf. The coefficient vectors A and C are calculated as follows:

$$A = 2a.r_1 - a, C = 2.r_2 \tag{8}$$

where r_1 and r_2 are random vectors uniformly distributed within the range [0, 1]. The exploration rate (a) decreases linearly from 2 to 0 over the course of iterations, as defined by the following equation:

$$a = 2 - t.(2/Max_{Iter}) \tag{9}$$

where t represents the current iteration, Max_{Iter} denotes the maximum allowed iterations for the optimization process. As a result, the random vector A is confined to the interval [-a, a]. This adaptive mechanism enables GWO to effectively balance exploration and exploitation during the search process.

Assuming that the wolves with α, β, and δ have better insight into the optimal solution space, the remaining wolves are compelled to adjust their positions in accordance with the leading wolves' movements.

$$D\alpha = |C1.X\alpha - X|,\ D\beta = |C2.X\beta - X|,\ D\delta = |C3.X\delta - X| \tag{10}$$

where Xα, Xβ, and Xδ are the best three solutions at a given iteration.

$$X1 = X\alpha - A1.(D\alpha),\ X2 = X\beta - A2.(D\beta),\ X3 = X\delta - A3.(D\delta) \tag{11}$$

$$X(t + 1) = (\ X1+ X2+ X3)/3 \tag{12}$$

Proposed Hybrid BAGWO Algorithm

The hybrid BAGWO algorithm leverages the complementary strengths of the GWO and BA. GWO's exploration capabilities are combined with BA's exploitation potential to enhance optimization performance. In the proposed approach, BA iteratively searches for promising solutions, and the top solutions are transferred to GWO for further refinement. This synergistic approach effectively explores the search space and identifies optimal solutions. Algorithm 1 outlines the pseudocode for the BAGWO algorithm.[16, 18]

Algorithm 1 Pseudo Code of BAGWO

```
initialize the BA population $X_i$ ($i$= 1,2,…,$n$) and $V_i$ define
pulse frequency $f_i$ at $X_i$
initialize pulse rates $r_i$, the loudness $A_i$
    while (t ≤ Max number. of iterations)
    for i= 1:n
    generate new solutions by adjusting frequency, and
    updating velocities and locations/solutions
    if (rand > $r_i$) then
      Select the best solution
      Generate a local solution around the best solution
    end if
    if (rand < A & fit(x) < fit($x^{t-1}$)) then
    Accept the new solution
    Increase $r_i$ and reduce $A_i$
    end if
    end for
    Rank the bats and find the current best
    end while
Initialize the **GWO** population $X_i$ ($i$= 1,2,…,$n$) and algorithm
parameters
   Evaluate the fitness of each search agent $fit_i$
   Initialize the first best solution as $X_\alpha$,
   Second best solution as $X_\beta$ and
   Third best solution as $X_\delta$
   initialize pulse rates ri, $a$, $A$ and $C$, t=0
   While ($t \le$ Max number. of iterations)
   for $i$=1: $n$
        Update the current search agent position
   End for
   Evaluate the fitness $fit_i$
   Update the coefficient vector $a$, $A$ and $C$
   If any better solution then
   update the best agents $X_\alpha$, $X_\beta$, $X_\delta$
   $t$=$t$+1
   End while
Return the first best agent $X_\alpha$ found so far.
```

Results and Discussion

Analysing the suggested algorithm's effectiveness is crucial now that the theoretical basis of the algorithm has been chosen. Therefore, we want to conduct tests to make sure the suggested method is effective. We have chosen to use the 10-fold cross-validation method in the current experiments to validate and record classification. To comprehensively evaluate the classifier's performance, we consider the following metrics:

- *Accuracy*: The percentage of all cases with accurate forecasts.
- *Precision*: The percentage of accurately predicted events that are really favourable.
- *Recall* (Sensitivity): The proportion of true positive cases that the model correctly categorises.
- *F1-score*: Precision and recall have a harmonic mean.
- *Specificity*: The proportion of actual negative cases that the model correctly classified as negative.
- *AUC-ROC*, or Area Under the Receiver Operating Characteristic curve, quantifies how well the model can discriminate between groups.

Using the entire dataset, we first evaluated the ML classifiers' performances in order to assess the effect of feature selection. This baseline study offers a standard by which to measure how well feature reduction works. Many classifiers were used, such as AdaBoost, Decision Table (DT), Fuzzy Unordered Rule Induction Algorithm (FURIA), C4.5, Random Forest (RF), K-nearest Neighbours

(K-NN), Multi-Layer Perceptron (MLP), Stochastic Gradient Descent (SGD), and Logistic Model Tree (LMT).

Performance of the Model Before Feature Selection

To establish a baseline for comparison, the full dataset was initially evaluated across nine ML classifiers. This analysis provides insights into the potential performance gains achievable through feature selection. So, prior to conducting the feature selection process, Table 1 presents a comparative analysis of various classification algorithms applied to the CKD dataset. RF exhibited superior performance with 99 % Accuracy, 99.3 % Precision, 98 % Recall, 98.3 % F1-score, and 98.8 % AUC. Other classifiers demonstrated strong performance, with accuracies exceeding 96 %. As the K-NN scored 98.5 % Accuracy, 96.8 % Precision, and 99.3 % Recall. The SGD obtained 98.3 % Accuracy, 95.5 % Precision, and 100 % Recall. The Accuracy, Precision, and Recall of the LMT algorithm were all 98 %, while the Accuracy of the MLP and DT was 97.5 %. The FURIA scored 97.3 % Accuracy, 96.6 % Precision, and 96 % Recall. With an Accuracy of 96.8 %, Precision of 96 %, and Recall of 96 %, the C4.5 produced the lowest results for this dataset.

Table 1. Classification performance before feature selection.

Classifier	Accuracy	Precision	Recall	F1-score	Specificity	AUC-ROC
MLP	0.975	0.943	0.993	0.958	0.964	0.979
SGD	0.983	0.955	1.000	0.970	0.972	0.986
K-NN	0.985	0.968	0.993	0.974	0.980	0.987
AdaBoost	0.963	0.953	0.947	0.935	0.972	0.959
DT	0.975	0.993	0.940	0.956	0.996	0.968
FURIA	0.973	0.966	0.960	0.952	0.980	0.970
C4.5	0.968	0.960	0.953	0.943	0.976	0.965
RF	0.990	0.993	0.980	0.983	0.996	0.988
LMT	0.980	0.961	0.987	0.966	0.976	0.981

The subsequent phase involved investigating the potential of nature-inspired algorithms, specifically the BA and GWO, for feature selection. The primary objective is to identify a subset of features that are most discriminative for the target classes while minimizing redundancy. This process aims to enhance classification accuracy by focusing on the most informative attributes within the dataset. Table 2 presents the features selected by the BA optimization algorithms in the first phase. The feature number was notably reduced from 24 to 10 features. Utilizing the BA optimization algorithm, the features urine specific gravity, serum albumin level, pus cell, plasma sodium, hemoglobin level, diabetes mellitus, coronary artery disease, appetite regulation, and anemia were recognized as the most significant for CKD prediction.

Table 2. Extracted features based on BA.

Algorithm	Selected Features	Features Number
BA	urine specific gravity serum albumin level pus cell plasma sodium haemoglobin level diabetes mellitus coronary artery disease appetite regulation pedal edema anemia	10

Table 3 presents a comparative analysis of classification algorithms applied to the CKD dataset using the features selected by the BA optimization algorithm in the first phase. RF consistently outperformed other classifiers, achieving an accuracy of 99.5 %, precision of 99.3 %, recall of 99.3 %, F1-score of 99.1 %, and AUC of 99.5 %. Other classifiers demonstrated strong performance, with accuracies exceeding 96.5 %. Notably, LMT achieved an accuracy of 98.8 % with a precision of 97.4 % and a recall of 99.3 %. FURIA and MLP also exhibited excellent performance, with accuracies exceeding 98 %. K-NN and C4.5 produced slightly lower results, with accuracies of 97.5 % and 96.8 %, respectively.

Table 3. Performance Assessment Using Selected Features Based on the BA.

Classifier	Accuracy	Precision	Recall	F1-score	Specificity	AUC-ROC
MLP	0.983	0.955	1.000	0.970	0.972	0.986
SGD	0.975	0.938	1.000	0.958	0.960	0.980
K-NN	0.975	0.949	0.987	0.957	0.968	0.977
AdaBoost	0.970	0.960	0.960	0.948	0.976	0.968
DT	0.978	0.986	0.953	0.960	0.992	0.973
FURIA	0.985	0.993	0.967	0.974	0.996	0.981
C4.5	0.965	0.966	0.940	0.938	0.980	0.960
RF	**0.995**	0.993	0.993	0.991	0.996	0.995
LMT	0.988	0.974	0.993	0.979	0.984	0.989

Table 4 presents the features selected by the GWO optimization algorithms in the second phase. The resulting feature number was reduced from 10 to 6 features. The BAGWO algorithm identified urine specific gravity, serum albumin level, haemoglobin level, diabetes mellitus, plasma sodium, and appetite regulation as the most significant predictors of CKD.

Table 4. Extracted Features Using BAGWO algorithm.

Algorithm	Selected Features	Features Number
BAGWO	urine specific gravity serum albumin level plasma sodium haemoglobin level diabetes mellitus appetite regulation	6

Table 5 presents an evaluation of the proposed BAGWO model's performance in early CKD detection and diagnosis. While RF consistently outperformed other classifiers, achieving a peak Accuracy of 99.5%, all models demonstrated reasonable performance. Notably, K-NN and FURIA with SGD achieved Accuracies of 99 % and 98 %, respectively. These results underscore the effectiveness of the proposed approach in enhancing CKD prediction. The superior performance of the proposed models positions them as valuable decision-support tools for the early diagnosis of CKD. By identifying key predictive features, these models can aid clinicians in making timely and informed decisions. The implementation of feature selection techniques has significantly enhanced the models' accuracy and efficiency, underscoring the importance of data pre-processing in developing robust predictive models for complex medical conditions.

Table 5. Performance Assessment Using Selected Features Based on the BAGWO.

Classifier	Accuracy	Precision	Recall	F1-score	Specificity	AUC-ROC
MLP	0.978	0.955	0.987	0.962	0.972	0.979
SGD	0.980	0.949	1.000	0.966	0.968	0.984
K-NN	0.990	0.980	0.993	0.983	0.988	0.991
AdaBoost	0.970	0.960	0.960	0.948	0.976	0.968
DT	0.978	0.986	0.953	0.960	0.992	0.973
FURIA	0.980	0.980	0.967	0.965	0.988	0.977
C4.5	0.965	0.966	0.940	0.938	0.980	0.960
RF	**0.995**	0.993	0.993	0.991	0.996	0.995
LMT	0.978	0.961	0.980	0.961	0.976	0.978

The proposed model demonstrates enhanced diagnostic capabilities for CKD, effectively predicting the disease using only 25 % of the original features. Table 6 provides a detailed performance evaluation, highlighting the model's superiority compared to previous studies using the UCI CKD dataset. While the model shows promise, limitations such as sample size and the absence of an independent validation set warrant further investigation to solidify its generalizability. A larger dataset would provide a more robust representation of the target population, increasing the confidence in the model's predictions. An independent validation set allows for a more objective evaluation of the model's performance.

Table 6. Comparison with other studies.

Reference	ML Classifier	Accuracy
[11]	RF	96%
[8]	Light GBM	99%
[9]	Gradient Boosting	99%
[6]	Support Vector Machine	99.33%
[7]	cost-sensitive AdaBoost	99.8%
[10]	NN	100%
[12]	Deep Learning NN	100%
Case under study	**RF**	**99.5%**

Conclusions

Millions of people worldwide are impacted by CKD, a serious global health issue. Given that kidney damage may be irreparable and that the condition frequently progresses without symptoms, early discovery and treatment are essential. In order to help with early CKD diagnosis and enhance patient outcomes, they created a prediction model. This study optimised CKD prediction by utilising these algorithms in light of the increasing interest in using ML for healthcare applications. To maximise model performance, nine ML classifiers were combined with a two-level feature selection method inspired by nature. This method reduced the feature set to only six properties, yet it greatly improved the classification accuracy. With an astounding accuracy of 99.5 %, RF was the best-performing classifier. This research can significantly improve patient outcomes and optimize healthcare resource allocation by facilitating early diagnosis and intervention.

Acknowledgements

This research was supported by the Public Authority for Applied Education and Training, Grant number BS-23-04.

References

1. Aminu K. Bello, et al., “An Update on the Global Disparities in Kidney Disease Burden and Care across World Countries and Regions,” *The Lancet Global Health* 12, no. 3 (2024): e382-e395.
2. Michel Burnier and Aikaterini Damianaki, “Hypertension as Cardiovascular Risk Factor in Chronic Kidney Disease,” *Circulation Research* 132, no. 8 (2023): 1050-1063.
3. Peter Rossing, Tine Willum Hansen, and Thomas Kümler, “Cardiovascular and Nonrenal Complications of Chronic Kidney Disease: Managing Risk,” *Diabetes, Obesity and Metabolism* Suppl 6 (2024): 13-21, https://doi.org/10.1111/dom.15747.
4. Wen-Tao Wu, et al, “Data Mining in Clinical Big Data: the Frequently Used Databases, Steps, and Methodological Models,” *Military Medical Research* 8 (2021): 1-12.
5. Lidong Wang and Cheryl Ann Alexander, “Big Data Analytics in Medical Engineering and Healthcare: Methods, Advances and Challenges,” *Journal of Medical Engineering & Technology* 44, no. 6 (2020): 267-283.
6. Debabrata Swain, et al, “A Robust Chronic Kidney Disease Classifier Using Machine Learning,” *Electronics* 12, no. 1 (2023): 212.
7. Sarah A. Ebiaredoh-Mienye, et al., “A Machine Learning Method with Filter-based Feature Selection for Improved Prediction of Chronic Kidney Disease,” *Bioengineering* 9, no. 8 (2022): 350.
8. Afia Farjana, et al., “Predicting Chronic Kidney Disease Using Machine Learning Algorithms,” *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, 2023.
9. Md Ariful Islam, Md Ziaul Hasan Majumder, and Md Alomgeer Hussein, “Chronic Kidney Disease Prediction Based on Machine Learning Algorithms,” *Journal of Pathology Informatics* 14 (2023): 100189.
10. Md Mehedi Hassan, et al., “A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records,” *Human-Centric Intelligent Systems* 3, no. 2 (2023): 92-104.
11. Chamandeep Kaur, et al., “Chronic Kidney Disease Prediction Using Machine Learning,” *Journal of Advances in Information Technology* 14, no. 2 (2023): 384-391.
12. Rahul Sawhney, et al., “A Comparative Assessment of Artificial Intelligence Models Used for Early Prediction and Evaluation of Chronic Kidney Disease,” *Decision Analytics Journal* 6 (2023): 100169.

[13] Manik Sharma and Prableen Kaur, "A Comprehensive Analysis of Nature-inspired Meta-heuristic Techniques for Feature Selection Problem," *Archives of Computational Methods in Engineering* 28 (2021): 1103-1127.

[14] Pratham Yashwante, et al., "Comparative Analysis of Meta-heuristic Feature Selection and Feature Extraction Approaches for Enhanced Chronic Kidney Disease Prediction," *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Vol. 2, IEEE, 2024.

[15] Aditya Khamparia, et al., "KDSAE: Chronic Kidney Disease Classification with Multimedia Data Learning Using Deep Stacked Autoencoder Network," *Multimedia Tools and Applications* 79 (2020): 35425-35440.

[16] Xin-She Yang, "Bat Algorithm for Multi-objective Optimisation," *International Journal of Bio-Inspired Computation* 3, no. 5 (2011): 267-274.

[17] Waqas Haider Bangyal, et al, "An Improved Bat Algorithm Based on novel Initialization Technique for Global Optimization Problem," *International Journal of Advanced Computer Science and Applications* 9, no. 7 (2018).

[18] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis, "Grey Wolf Optimizer," *Advances in Engineering Software* 69 (2014): 46-61.

[19] Yudong Zhang, Shuihua Wang, and Genlin Ji, "A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications," *Mathematical Problems in Engineering* 2015, no. 1 (2015): 931256.

About the Authors

Waheeda I. **Almayyan** is a Professor of Computer Science and Information Systems at the College of Business Studies, Public Authority for Applied Education and Training, Kuwait. Holding a Ph.D. in Artificial Intelligence with a focus on biometric authentication, and Master's and Bachelor's degrees in Computer Science, Dr. Almayyan's research interests span data mining and systems usability.
https://orcid.org/0009-0003-8272-6293

Bareeq A. **AlGhannam** is an Associate Professor in the Computer Science and Information Systems Department, College of Business Studies, the Public Authority for Applied Education and Training in Kuwait. Dr. AlGhannam has a Bachelor and Master of Science in Computer Engineering with a Ph.D. in Software Engineering focused on the realm of Stakeholder Collaboration within software requirements collection. Currently she is conducting various research in Systems Usability and Data Mining.
https://orcid.org/0000-0002-8724-7414