Artificial intelligence for early-stage detection of chronic kidney disease

Mamatha B.[1], **Sujatha P. Terdal**[2]
[1]Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), CMR Technical Campus, Hyderabad, India
[2]Computer Science and Engineering, PDA College of Engineering Gulbarga, Kalaburagi, India

Article Info

Article history:

Received Jan 3, 2024
Revised Mar 26, 2024
Accepted Apr 16, 2024

Keywords:

Apache spark
Artificial intelligence
Big data analysis
Chronic kidney disease
Enhanced hybrid network

ABSTRACT

Early-stage detection of chronic kidney disease (CKD) is crucial in research to enable timely intervention, enhance understanding of disease progression, reduce healthcare costs and support public health initiatives. The traditional approaches on early-stage chronic kidney disease detection often suffer from slow convergence and not integrate advanced technologies, impacting their effectiveness. Additionally, security and privacy concerns related to patient data are ineffectively addressed. To overcome these issues, this research incorporates novel optimized artificial intelligence-based approaches. The main aim is to enhance detection process through enhanced hybrid mud ring network (EHMRN), a novel detection technique combining light gradient boosting machine and MobileNet, involving extensive data collection, including a large dataset of 100,000 instances. The introduced network is optimized through the mud ring optimization to attain enhanced performance. Incorporating spark ensures secure cloud-based storage, enhancing privacy and compliance with healthcare data regulations. This approach represents a significant advancement in primary stage detection more effectively and promptly. The results show that the introduced approach outperforms traditional approaches in terms of accuracy (99.96%), F1-score (99.91%), precision (100%), specificity (99.98%), recall (100%) and execution time (0.09s).

Corresponding Author:

Mamatha B.
Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), CMR Technical Campus
Hyderabad, Andhra Pradesh-501401, India
Email: mamatha.789@gmail.com

1. INTRODUCTION

Chronic kidney disease (CKD) poses a significant global health challenge, characterized by a gradual decline in kidney function. Despite its asymptomatic nature, CKD progress silently to advanced stages, leading to severe complications and increased healthcare burdens. The imperative for early-stage CKD detection is underscored by its potential to substantially improve patient outcomes through timely intervention and management [1]. Recognizing the importance of this issue, this research seeks to address existing challenges in CKD detection by employing a novel big data analytics approach [2], [3].

Traditional methods for detecting CKD encounter significant challenges when faced with large-scale datasets, managing missing data, and selecting optimal features. These limitations hamper their effectiveness in early detection. Recent advancements in machine learning (ML) and deep learning (DL) have introduced various techniques tailored for early-stage detection of chronic kidney disease (ESDCKD).

Deep neural network (DNN) and an adaptive moment estimation optimization function is developed in [4] to predict early-stage CKD. To enhance interpretability, predictions of the DNN-CKD model are explained using the local interpretable model-agnostic explainer (LIME). The model is trained on diagnostic patient data using a five-layered DNN with three hidden layers. However, this approach is susceptible to overfit and require substantial computational resources. By adding measurable features, convolutional neural network (CNN) model [5] is developed for improved CKD detection in ultrasound images. Using a mask regional CNN, organ segmentation and feature extraction, such as kidney length and echogenicity ratio, were performed. The ResNet18 model classified images into CKD and non-CKD groups, with and without measurable feature input. But this technique is computationally intensive and attains a minimum convergence. A method in [6] aims to create a predictive model for CKD heart disease data using various open-source Python modules. By comparing with algorithms like k-nearest neighbor (KNNs) and recurrent neural network (RNN), the goal is to achieve high prediction and precision in machine learning methods. Although, this approach face issues with vanishing/exploding gradient problems, making them challenging to train effectively. The algorithm in [7] utilizes a two-layer architecture with real-time data. It employs artificial neural networks (ANN) to learn features and extreme gradient boosting (XGB) for predictions which struggle with noisy data and exhibit slow training on large datasets. New CKD detection model including preprocessing, segmentation, feature extraction and classification is developed in [8]. The output features are fed into long short-term memory (LSTM) for improved accuracy. But LSTM is associated with computational complexity and a risk of overfitting. Addressing these drawbacks is essential for optimizing the effectiveness of these techniques in early-stage CKD detection.

Addressing the existing challenges in machine learning entails overcoming hurdles such as overfitting, resource-intensive requirements, difficulties with vanishing/exploding gradients, handling noisy data effectively, improving training efficiency on large datasets, enhancing scalability, ensuring interpretability, simplifying complexity, minimizing sensitivity to hyperparameter tuning, and mitigating the risk of overfitting. Despite the advancements in CKD detection techniques, several unsolved problems persist, including the need for more accurate and efficient detection methods, especially in early-stage CKD. To address these challenges, this research proposes a novel approach that leverages big data analytics to augment the dataset size for comprehensive analysis. This approach is a significant departure from traditional methods and contributes to the field by offering a novel perspective on CKD detection [9], [10].

In this research, a comprehensive methodology is introduced aimed at advancing early-stage CKD detection by fusing artificial intelligence (AI) based techniques. Specifically, enhanced light gradient boosting machine (ELightGBM) and MobileNet (MN) is utilized and optimized through mud ring optimization algorithm (MROA), to enhance accuracy and efficiency compared to traditional methods. Additionally, generative adversarial network (GAN) is employed for big data analytics, expanding the dataset size and contributing to a more thorough analysis of CKD.

The contributions of this research are as follows: i) The use of generative adversarial networks (GANs) in the big data analytics process aims to replicate the original data's characteristics, thereby expanding the dataset size for comprehensive analysis. This innovative approach to data augmentation represents a significant contribution to chronic kidney disease detection, as it introduces a novel method not previously explored in existing literature. The preprocessing of big data using techniques like part mean imputation (PMI) and improved binning (IB) further enhances the quality of the dataset; ii) The integration of the enhanced hybrid mud ring network (EHMRN), comprising ELightGBM and MN optimized through MROA, significantly improves accuracy and efficiency compared to conventional methods. This advancement addresses convergence challenges and provides a more robust solution with enhanced training for ESDCKD detection; iii) The utilization of MROA for tuning neural network parameters is driven by its exceptional adaptability to complex parameter spaces. This algorithm ensures rapid convergence and superior accuracy in CKD detection, surpassing alternative tuning methodologies; iv) The adoption of the modified energy valley optimizer (MEVO) is based on its outstanding capability to extract essential features efficiently for optimizing CKD prediction by limiting the time complexity. MEVO outperforms other algorithms by effectively selecting key data inputs, thereby enhancing detection accuracy; and v) Incorporating spark for secure cloud-based storage addresses critical privacy concerns and ensures compliance with healthcare data regulations, demonstrating a commitment to data security and privacy in CKD detection efforts.

The rest of the research is organized as follows: Section 2 delves into an in-depth examination of recent literature, thoroughly surveying existing research findings and developments in the field. Following this, section 3 introduces the topology employed in this research, offering a detailed description of the framework or structure utilized for analysis. Subsequently, section 4 meticulously presents the outcomes derived from the implemented methodology, providing a comprehensive overview of the results obtained. Lastly, section 5 encapsulates the study with a concluding discussion, summarizing key findings and insights drawn from the research process.

2. REVIEW

This section provides a concise review of recent approaches and limitations in CKD and ESDCKD. Modified extreme-random forest (ME-RF) with eXtreme gradient boosting (XGBoost) integration was applied by Rajeashwari and Arunesh [11] for classification, incorporating increased decision tree (DT) leaves to enhance adaptability. Evaluation on datasets with varied deep CNN (Deep CNN) configurations revealed high computational demands and limited interpretability as practical constraints. Rahman *et al.* [12] examines eight ensemble ML (EML) methods for early CKD diagnosis on ML datasets. It addresses missing data with multiple imputations by chain equation (MICE) imputation and handles data imbalance with borderline support vector machine with synthetic minority over-sampling (SVMSMOTE). Additionally, it employs feature selection techniques and hyperparameter tuning to enhance classifier performance and efficiency. This approach was unable to identify the stage of the disease. Kidney fibrosis in patients with CKD was predicted by Ge *et al.* [13] using radiomics of two-dimensional ultrasound and sound touch elastography (STE) images in grouping with clinical attributes. Yet, this method showed inaccuracies in its classification. Islam *et al.* [14] explored hybrid machine learning (HML)'s potential for early CKD diagnosis, emphasizing the importance of timely intervention. Employing predictive modeling and testing twelve classifiers, the study identifies XGBoost as the most effective, showcasing advancements in ML for accurate kidney disease prediction. The model faced challenges due to the insufficient sophistication and representativeness of the available data. Support vector machine (SVM) was utilized as a classifier by Swain *et al.* [15] to develop a predictive model, with public data being leveraged for forecasting CKD. Data preprocessing and the chi-squared test for feature extraction were employed in the pursuit of a robust predictive model. Yet, this method experienced lower accuracy rates and reduced effectiveness.

A novel ensemble deep learning (EDL) approach was introduced by Alsekait *et al.* [16] for detecting early CKD, incorporating various feature selection methods to identify optimal features. The impact of these features on CKD from a medical perspective was studied and the EDL models with the SVM as the meta learner model. But the suggested algorithm experiences elevated computational complexity. The Ebola deep wavelet extreme learning machine (EDWELM) learning developed by Reddy and Vydeki [17] for accurate early CKD and non-CKD classification, involving data preprocessing and feature selection using the darts battle game optimizer. The last step encompassed the detection process, crucial in data mining for characterizing data classes. This approach demonstrated improved time consumption, a weak convergence rate and training. An innovative fusion DL model was introduced by Rao *et al.* [18], combining a graph neural network and a tabular data learning (GNN-TDL) approach for detecting CKD evolution early, leveraging the strengths of both data representations. This approach faced increased computational demands and reduced model interpretability. A ML model for early CKD detection was developed by Pal *et al.* [19], involving applying baseline detectors to categorical and non-categorical attributes and enhancing outcomes by combining outputs through a majority vote. However, this approach shows weak prediction performance. A modified version of weighted mean of vectors (INFO) called mINFO developed by Houssein and Sayed [20] addresses challenges like local optima and slow convergence. Developed with opposition-based learning (OBL) and dynamic candidate solution (DCS) strategies, mINFO enhances local search and overcomes premature convergence issues while effectively managing CKD. However, mINFO faced limitations in effectively handling complex variations within CKD datasets, impacting its performance in certain scenarios. Self-attention convolutional neural network (SACNN) optimized with season optimization algorithm (SOA) called SACNN-SOA was presented by Alikhan *et al.* [21] in a smart medical big data healthcare scheme using internet of things (IoT) and cloud computing. SACNN was used without disclosing optimization systems; instead, optimization was carried out using the SOA. This model encounters increased computational complexity.

2.1. Problem statement

The existing research faces challenges. These challenges include diagnostic inaccuracy and increased execution time, necessitating improved diagnostic methods, faster research processes, and more effective therapeutic strategies to combat CKD's rising prevalence and address critical healthcare challenges. These limitations highlight the need for innovative approaches and solutions to overcome the obstacles in CKD research and healthcare interventions.

3. INTRODUCED TOPOLOGY

The introduced methodology is comprised of comprehensive four-phase approach to ESDCKD. The first stage involves a meticulous analysis of big data using GAN, delving into substantial datasets to extract valuable insights. Following this, a preprocessing step is implemented using PMI and IB to eliminate noise and address missing values, ensuring the integrity and accuracy of the data. The subsequent stage involves the selection of pertinent features using MEVO, a crucial step in enhancing the predictive capabilities of the model. Then, detection of the likelihood of CKD is then performed using EHMRN based on the refined data. Finally, the outcomes, indicating the probability of CKD occurrence or absence are securely stored in the Spark cloud, ensuring robust data management and accessibility. This methodological framework in Figure 1 underscores the significance of each stage, from data analysis to feature selection, in constructing a robust approach for early CKD detection. The secure storage of results in the spark not only facilitates efficient data management but also ensures the confidentiality and accessibility of critical health-related information.

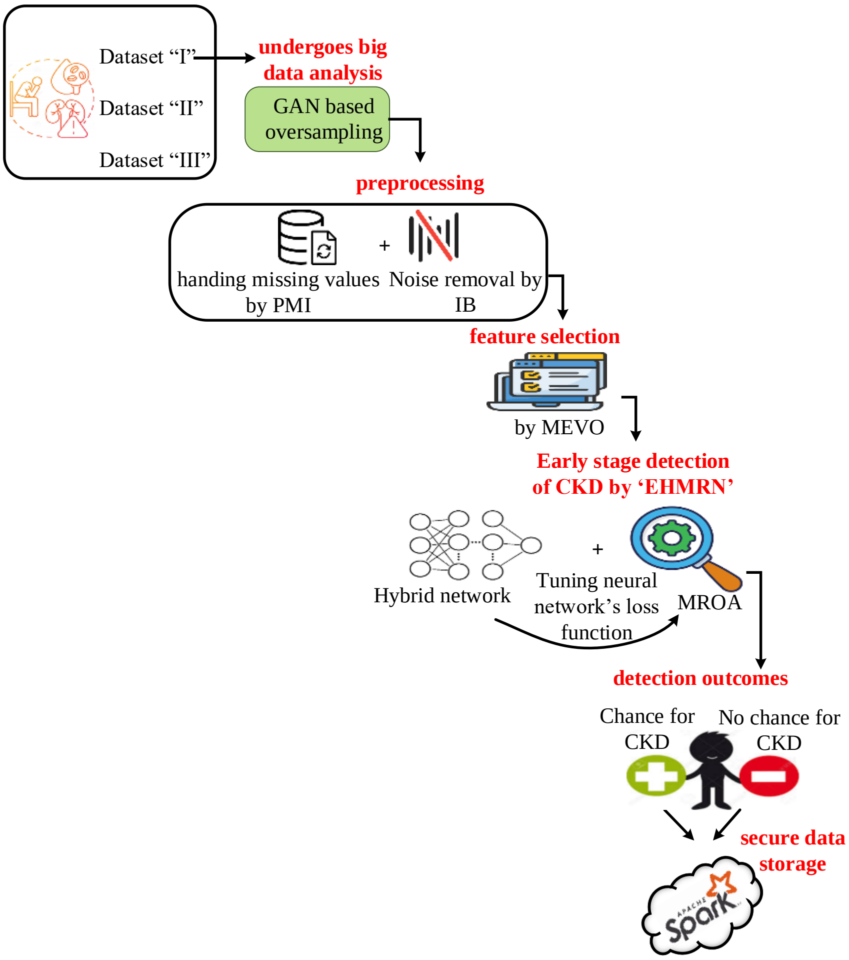


Figure 1. Architecture of introduced model

3.1. Input data collection

The data utilized as input for ESDCKD is derived from three distinct datasets: CKD (dataset "I") [22], CKD from Kaggle (dataset "II") [23] and CKD–explored (dataset "III") [24]. Dataset I, which originated from the UCI ML Repository, serves the purpose of predicting CKD. This dataset is typically collected from hospitals over a span of nearly two months, which encompasses 400 unique instances, each defined by 25 distinct features. Moving on, dataset II is specifically designed to ascertain whether a patient is afflicted with CKD or not, relying on a comprehensive set of diagnostic measurements incorporated within the dataset. Notably, this dataset integrates an array of medical predictor variables alongside a singular target variable. Meanwhile, dataset III is comprised of 24 features and one target variable, contributing to the comprehensive understanding of CKD.

3.2. Big data analysis

To further bolster the data collection process, an extensive dataset has been generated using AI based approaches, drawing its foundation from the initial dataset I. The dataset of 400 instances is expanded to 100,000 instances for big data analysis through a process called oversampling. Oversampling using GAN [25] involves creating additional synthetic instances in the dataset to rebalance the data and ensure that the model is trained on a more comprehensive and representative sample. This technique is commonly employed to address issues related to class imbalance, where certain classes are underrepresented in the dataset. The generator combines the input noise z and b to form a joint hidden representation. Similarly, a and b are fed into the discriminator. Likewise, the discriminator takes a and b as input. The oversampling learning objective of a GAN is represented by (1), where both the discriminator and generator now depend on b as an input.

$$GAN_{loss}(a,b,\theta,\varphi) = \theta_{min}\ \varphi_{max}\left(e_{a\_Pa}[log\ K_{\phi}\ (a,b)] + e_{c\_Pc}[log\ K_{\phi}\ (M_{\theta}(a,b),c))]\right) \tag{1}$$

After the training phase, when generating an observation for a class b, the process involves first sampling c from the distribution $QC(c)$. Subsequently, both the sampled c and the specified b are fed into the conditional generator to produce the resulting output $a^{'} = \Psi\theta(c,b)$. By artificially enlarging the instances of the minority class, the dataset becomes more balanced, thus enabling the approach to absorb the patterns and relations within the data well. This expanded dataset boasts a substantial compilation of 100,000 instances, accompanied by a diverse range of additional variables. Subsequently, all of this data undergoes a preprocessing stage.

3.3. Preprocessing phase

The input data from the aforementioned datasets contains both missing values and noisy data. To address this issue, the current research employs a preprocessing technique that involves PMI and IB. This method aims to effectively handle missing values by replacing them with the mean of the corresponding feature and it also involves grouping continuous data into intervals (or 'bins') for easier analysis and interpretation.

3.3.1. Part-mean imputation

In Part-mean imputation (PMI) [26], the missing values of the response variable in a specific part of the CKD datasets are substituted by the mean of the non-missing response parameters within that part. The missing response value mb_j is attained from the factorial portion $\hat{b}PMI_c = \frac{\sum_{\forall j\neq c} b_j}{3}, c \in j = \{1,2,3,4\}$, middle portion $\hat{b}PMI_c = \frac{\sum_{\forall j\neq c} b_j}{4}, c \in j = \{5,6,7,8,9\}$, axial portion $\hat{b}PMI_c = \frac{\sum_{\forall j\neq c} b_j}{3}, c \in j = \{10,11,12,13\}$, using PMI as in (2). The PMI (2) is applied to determine the parameter of the missing response at the design point within the factorial portion F_4 in relation to b_{13}.

$$\hat{b}_{PMI} = \frac{b_{10}+b_{11}+b_{12}}{3} \tag{2}$$

3.3.2. Improved binning

The other preprocessing approach is based on the IB [27] of numerical variables, wherein distinct numerical parameters are converted into ranges and classes like low (<0,4>), medium (<5,6>), high (<5,8>) and very high (<9,10>) are used. The discrete scale of 0–10 in the benchmark scores is mapped into categorical labels using this approach. This development of the IB process stemmed from interpreting and comprehending these categories based on the perspective of business expertise. By grouping numerical values into meaningful categories with the input of domain experts, IB smooth out noise. Following the preprocessing stage, the finest features are chosen from the preprocessed data using the MEVO method.

3.4. Feature selection phase

The modified form of energy valley optimizer (EVO) [28] algorithm known as MEVO is used for selecting the finest features from the preprocessed output because this algorithm excels in identifying crucial features within a dataset, handling diverse data structures and providing interpretability features, ensuring its position as a robust tool for complex data analysis. This modification reduces time complexity by implementing efficient search strategies that efficiently explore the solution space, leading to faster convergence and reduced computational overhead. Within the MEVO framework, the population size is divided into two distinct segments which enhances the algorithm's convergence speed and improves the effectiveness of the algorithm. Initially, proportion of the injected population ratio (25%, 50%, 75%, and 100%) are done by the modified technique. Features I_{MI} with high mutual information (MI) values are ensured inclusion in the initial population, guaranteeing their significance for classification which is determined in (3), where the binary representation of the i^{th} feature in the initial population is represented by I_p and a random number within the range of [0, 1] is denoted as r. The second portion indicates the remaining population (1-injected population ratio), which is randomly initialized using (4). The fitness value for choosing the finest features is determined using (5).

$$I_p = \begin{cases} 1, ifr < normalized \rightleftarrows I_{MI} \\ 0, ifr \geq normalized I_{MI} \end{cases} \tag{3}$$

$$I_p = \begin{cases} 1, ifr > 0.5 \\ 0, ifr \leq 0.5 \end{cases} \tag{4}$$

$$O = a \cdot (|fpr - fnr|) + b \cdot \frac{|F|}{|T|} \tag{5}$$

where, the weight of each objective is represented by parameters a and b, both of which fall between zero and one (with b being derived as $1 - a$). The number of selected and total features are denoted by F and T. fnr and fpr indicate the false negative rate (FNR) and false positive rate (FPR), respectively. In this research work, a is predetermined as 0.99 and b is set to 0.01. The utilization of (5) facilitates the selection of the best features, which are subsequently applied for ESDCKD.

3.5. Detection phase

The introduced hybrid network [29], [30] is designed to accept a specific set of chosen features as its input. These features play a crucial role in enabling the network to effectively differentiate instances of CKD with precision. By leveraging these carefully selected features, the hybrid network endeavors to achieve accurate identification and classification of CKD cases. Notably, the predictions generated by the MN component of the hybrid network serve as input labels for the training process of the ELightGBM model, contributing to the model's learning and predictive capabilities.

3.5.1. MobileNet detection

MobileNet classifies medical images into categories indicating the presence or absence of CKD. The output of MN O_{MN} represents the initial predictions. Following the initial predictions made by the MN component, these outcomes serve as essential input labels for the subsequent stage involving the ELightGBM model. This model, being an integral part of the process, undertakes further refinement of the detections based on the provided input labels. By leveraging this refined information, the ELightGBM model then generates the final likelihood estimates for CKD, which encapsulate a more comprehensive and accurate assessment of the presence of CKD within the analyzed data.

3.5.2. Enhanced LightGBM model training

The MN predictions O_{MN} serve as input labels for training the LightGBM model. The objective function (OF) for ELightGBM O_i and its components remain consistent with the standalone ELightGBM in (6), where $i = 1,2,..n$ signifies the overall samples of training, L signifies the loss function, y_i is the true label of the i^{th} training sample, $L(y_i, F(x_i))$ is the loss incurred for predicting when the true label is y_i. The ELightGBM model's loss function is improved by incorporating a dynamic loss mechanism to address the vanishing gradient problem, resulting in improved training convergence performance. This dynamic loss is formally defined in the accompanying (7).

$$O_i = \sum_{i=1}^{n} L(y_i, F(x_i)) \tag{6}$$

$$L(x, \beta, y) = \begin{cases} 0.5(x/y)^2 \beta = 2 \\ log((x/y)^2 + 1)\, \beta = 0 \\ 1 - exp(-0.5(x/y)^2)\, \beta = -\infty \\ L(x, \beta, y) others \end{cases} \tag{7}$$

When considering a residual value represented by x, the hyperparameter β plays a crucial role in adjusting various components within the loss function. Additionally, the coordination parameter y is employed to fine-tune the bending scale of the loss function at $x = 0$, thereby determining its appropriateness for the introduced gradient-detection method. The primary objective in the training process is to minimize the incurred loss across all samples. This training loss is subsequently reduced by incorporating a MROA known for its effective exploration capability and faster convergence rate. To achieve optimal detection performance, the hybrid network model's hyperparameter $\beta$and loss function L in (7) need to be optimized. Therefore, these tuning of parameters serve as the fitness evaluation function for the mud ring optimization algorithm (MROA).

3.6. Tuning of loss function and hyperparameter

The MROA is favored over traditional algorithms for its exceptional performance in exploration and convergence rates, offering adaptability in complex search spaces and efficient solutions for diverse optimization problems. The OF for achieving improved detection outcomes is derived by the MROA approach. A numerical value is specified to represent the improved effectiveness of potential solutions. In this research work, the OF in (8) is considered as the tuning of β and L of the hybrid network (from (7)). This OF mentioned in (8) is achieved using the exploration and exploitation stages of MROA [31]. The overall process of MROA is depicted in Figure 2.

$$OF(O_i) = tuning\{L,\beta\} = \frac{\text{misclassified}samples}{overallsamples} \times 100 \tag{8}$$

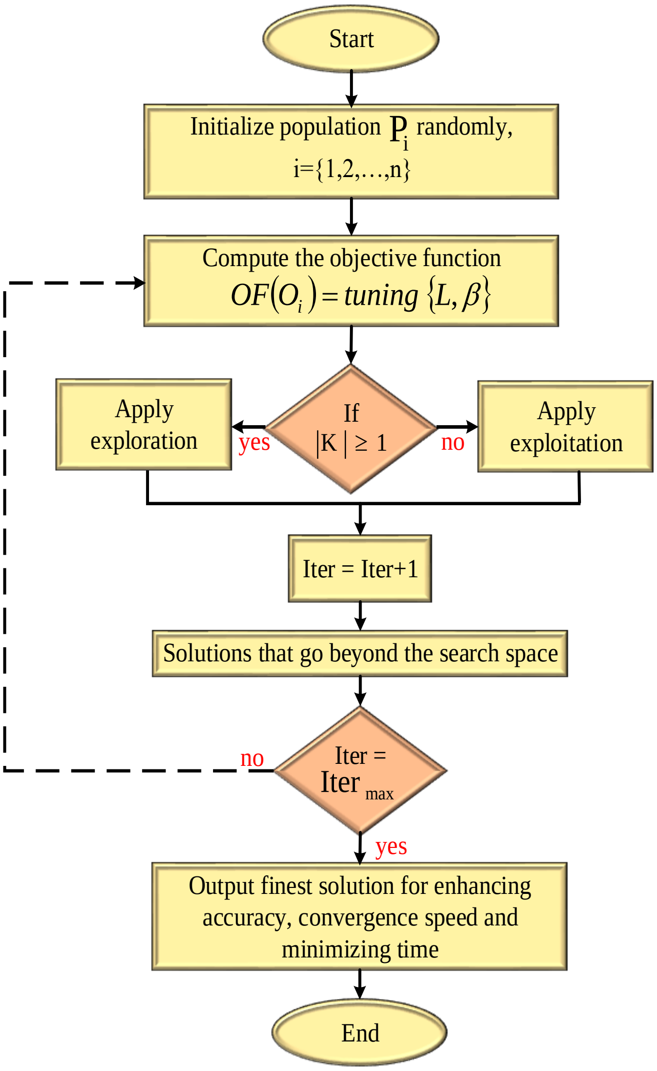

Start
Initialize population P_i randomly, i={1,2,…,n}
Compute the objective function $OF(O_i) = tuning\{L,\beta\}$
If |K| ≥ 1
yes
no
Apply exploration
Apply exploitation
Iter = Iter+1
Solutions that go beyond the search space
Iter = Iter max
no
yes
Output finest solution for enhancing accuracy, convergence speed and minimizing time
End

Figure 2. Flowchart of MROA

The EHMRN, which integrates ELightGBM and MN optimized through MROA, enhances accuracy by leveraging the strengths of both models and optimizing their performance through advanced optimization techniques. ELightGBM is known for its efficient gradient boosting capabilities, while MN provides deep learning capabilities for complex pattern recognition. The MROA optimization further refines the model parameters, improving convergence and reducing overfitting, resulting in enhanced accuracy. This integration and optimization process synergistically improve the model's ability to capture intricate patterns and make accurate predictions, making EHMRN a powerful tool for achieving high accuracy in various machine learning tasks.

Hence, the introduced AI based approaches has accurately executed ESDCKD. The methodology has demonstrated its capability to identify and address early signs of kidney disease, showcasing its proficiency in contributing to the advancement of diagnostic processes for renal health. Subsequently, the collected data undergo a process of aggregation and are securely stored within the spark cloud infrastructure. This entails the systematic compilation and organization of the acquired information, ensuring its integrity and accessibility within the spark cloud environment. The storage mechanism employed facilitates efficient data management, enabling seamless retrieval and analysis when needed. This step plays a pivotal role in establishing a robust and scalable data repository, contributing to the overall efficacy and reliability of the information storage system.

4. SIMULATION OUTCOMES

This section provides an extensive explanation of the results generated by the methodology introduced. The methodology was executed utilizing Apache Spark with PySpark in the Python programming language. The evaluation of this model's effectiveness is conducted by comparing it to several existing approaches. This comparative analysis offers valuable insights into the strengths and contributions of the introduced model.

Table 1 outlines key hyperparameters for training introduced neural network model. It specifies 100 epochs for overall training duration, up to 3 hidden layers using Rectified linear unit (ReLU) activation functions. The optimization is handled by the MROA optimizer, with a batch size of 16 and a learning rate set at 0.001. These parameters collectively guide the model's training process, influencing its performance and ability to learn from the data.

Table 1. Experimental settings of the introduced approach

Hyperparameters	Ranges
Overall training epochs	100
Overall hidden layers	3
Activation function	ReLU
Optimizer	MROA
Batch size	16
Learning rate	0.001

Figure 3 presents the confusion matrices, providing a detailed representation of the model's performance. Notably, the EHMRN exhibits a commendable level of accuracy by correctly identifying all instances of genuine true positive events (7,531 for dataset “I”, 19 for both datasets “II” and “III”) and true negative events (7,469 for dataset “I”, 40 for both datasets “II” and “III”). This evaluation through the confusion matrices offers a comprehensive insight into the model's discriminative capabilities, highlighting its effectiveness in distinguishing between positive and negative events in the analyzed datasets “I”, “II”, and “III”, as shown in Figures 3(a) dataset I, 3(b) dataset II, and 3(c) dataset III.

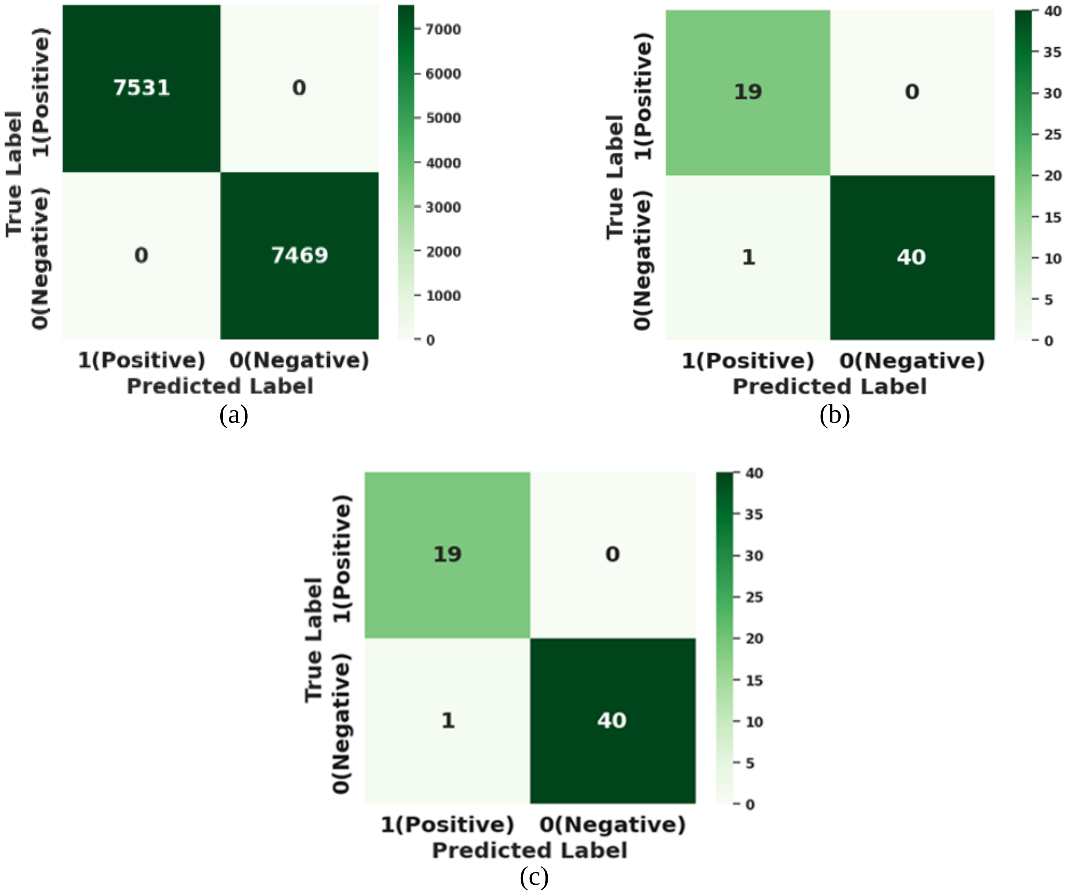
The absolute correlations between the class label and various features of datasets “I”, “II”, and “III” are elucidated in Figures 4(a)-(c). Noteworthy correlations include positive associations with blood pressure, specific gravity, albumin, sugar, blood urea, serum creatinine, blood glucose random and sodium, while hemoglobin, potassium, white blood cell count and red blood cell count exhibit negative correlations. This visual depiction of data relationships, showcased in Figure 4, allows for an intuitive examination of the intricate connections between features, offering valuable insights into their interdependencies within the dataset. The accuracy and loss curves, tracked across iterations in Figure, showcase the model's learning dynamics. In this context, the introduced approach excels by consistently achieving a superior steady-state accuracy value in Figures 5(a), 5(c) and 5(e) indicating its sustained high-level performance in making correct predictions. The corresponding loss curve in Figures 5(b), 5(d) and 5(f) underscores the model's effective error minimization, highlighting its efficiency in learning and accurate prediction.

Figures 6(a)-(e) showcases a comparative analysis of various AI based techniques for detecting CKD using accuracy, F1-score, precision, recall and specificity. Each technique is evaluated with respect to the key evaluation metrics, including accuracy, F1-score, precision, recall and specificity. EHMRN consistently outperforms established methods like deep convolutional neural network (Deep CNN) [11], ensemble machine learning (EML) [12], hybrid machine learning (HML) [14], support vector machine (SVM) [15], ensemble deep learning (EDL) [16], Ebola deep wavelet extreme learning machine (EDWELM) [17], graph neural network and a tabular data learning (GNN-TDL) [18], SACNN-SOA [21], and mINFO [20]. EML, EDL and EDWELM also demonstrate exceptional performance, consistently scoring above 99%

in all metrics. SVM and HML exhibit strong results with accuracy and precision and recall around at 99%. Deep CNN achieves high accuracy at 98.50% and well-balanced F1-score, precision, recall and specificity. GNN-TDL and SACNN-SOA show slightly lower scores, with GNN-TDL having an accuracy of 95.089% and SACNN-SOA achieving 99% accuracy but with a lower F1-score. mINFO performs well in accuracy 99.34% and precision 99.55% but has a comparatively lower F1-score 91.02% and recall 99.29%. Notably, the introduced approach, EHMRN, attains exceptional scores, achieving above 99.9% in all metrics, signifying its outstanding predictive capabilities. The uniform excellence across diverse metrics establishes EHMRN as a benchmark for accurate and reliable CKD prediction. Its optimal trade-off between sensitivity and specificity is particularly noteworthy, crucial for minimizing both false positive rates and false negative rates in medical applications. This adaptability underscores the robustness and versatility of the introduced technique beyond the scope of the current analysis.

Figure 7 outlines receiver operating characteristic (ROC) values for various AI based approaches, offering insights into their discriminatory performance in a binary classification task, potentially related to ESDCKD. EML displays robust discriminative ability with a ROC value of 0.9888, while SVM excels with 0.9932. EDWELM showcases near-perfect discrimination at 0.9983. An unspecified machine learning achieves a moderate ROC value of 0.77. SACNN-SOA exhibits strong discriminatory capabilities, achieving a ROC value of 0.97. Notably, the introduced approach, stands out with a perfect ROC value of 1, indicating flawless discriminatory performance and optimal classification in the evaluated task.

The precision-recall curve for the EHMRN in Figure 8 shows that it has high precision and recall at all thresholds, indicating that it is able to accurately identify both positive and negative cases. Figure 8 unequivocally demonstrates that the EHMRN surpasses all conventional algorithms in terms of accuracy. This advantage arises from integrating MROA with hybrid network, strategically avoiding local optima pitfalls. By navigating through the intricacies of the solution space, this combined approach successfully identifies the global optimum solution. The outcome is a substantially elevated accuracy rate, establishing EHMRN as a noteworthy advancement in achieving more accurate and dependable results compared to traditional algorithms.

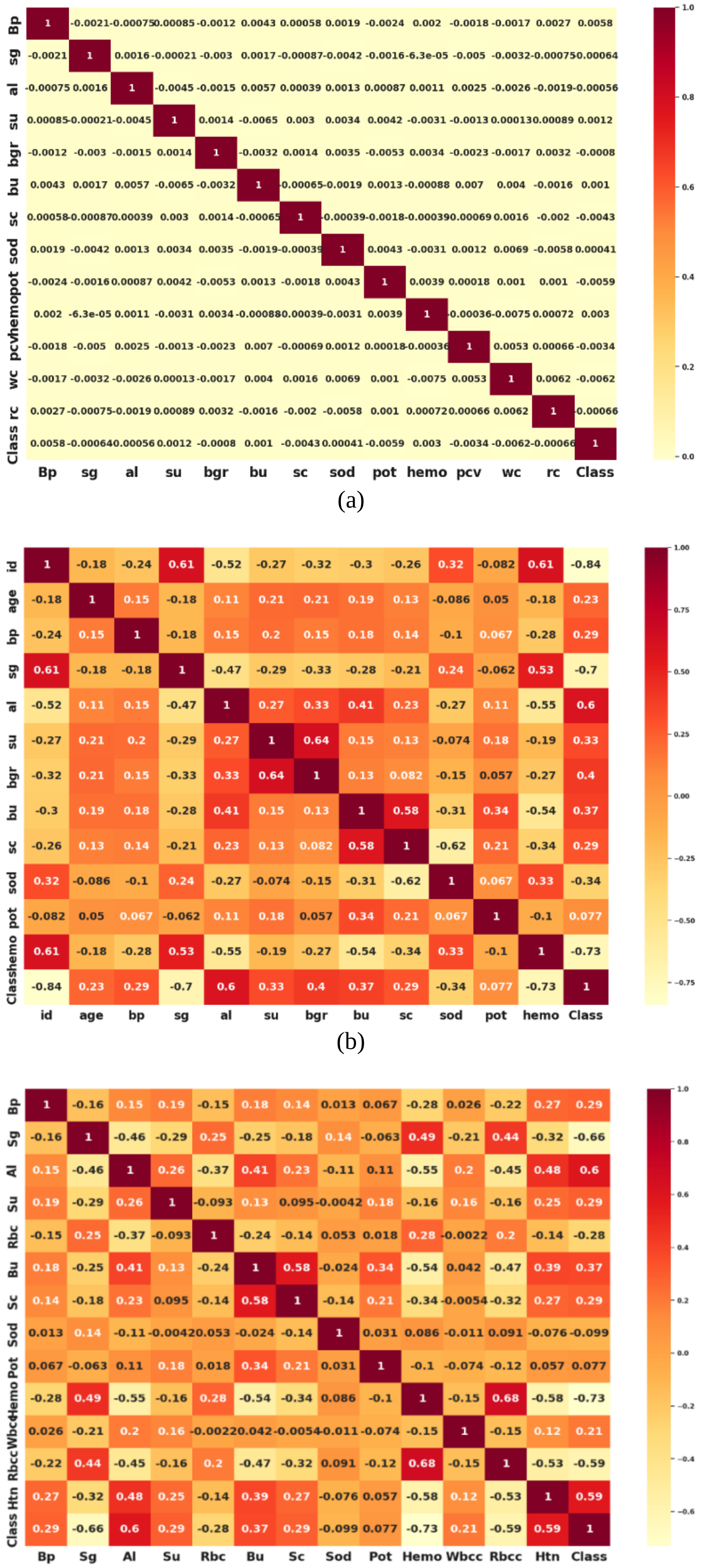


(a)

(b)

(c)

Figure 3. Confusion matrices of datasets (a) “I”, (b) “II”, and (c) “III”

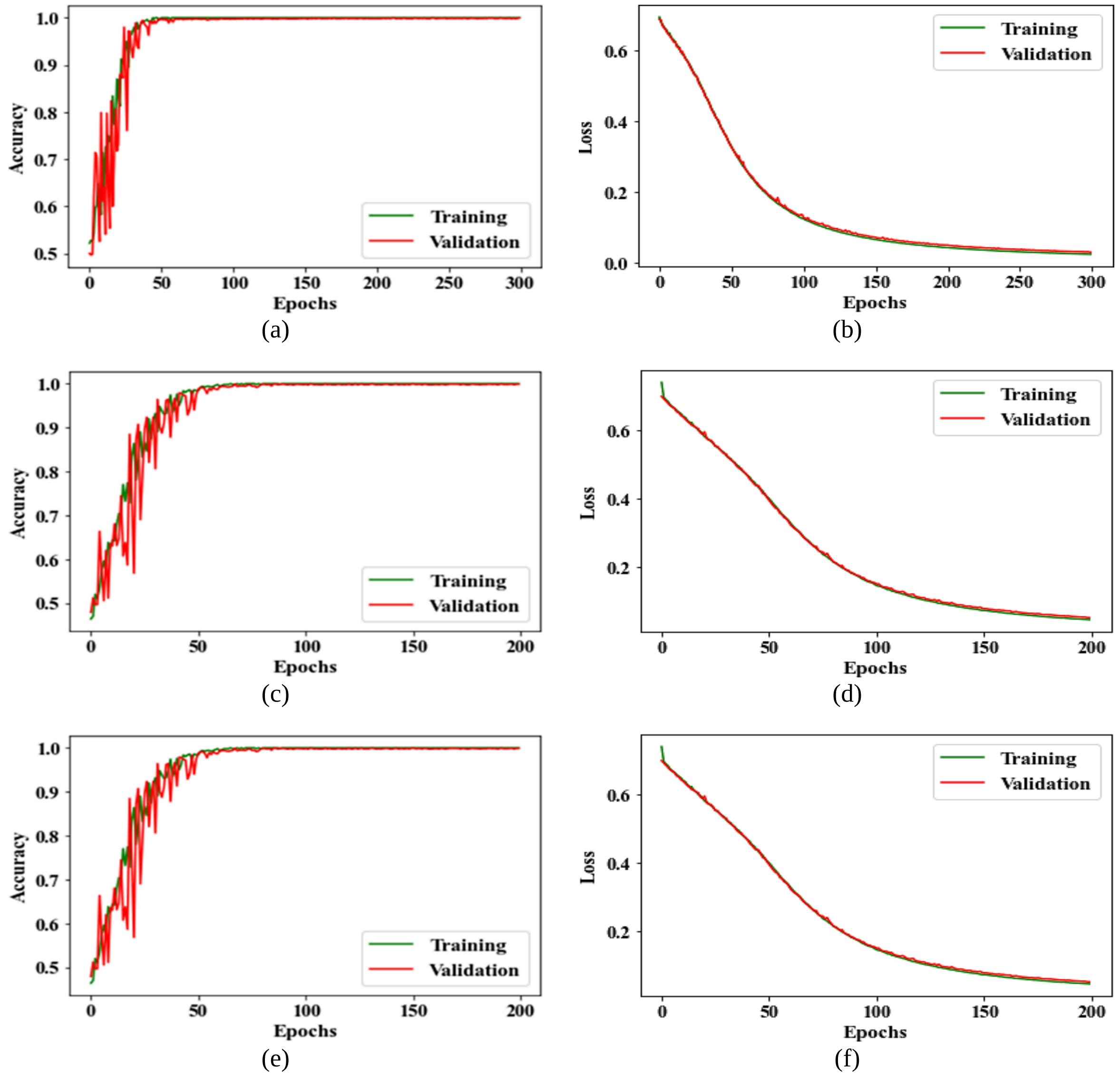


(a)

(b)

(c)

Figure 4. Correlation matrices of datasets (a) “I”, (b) “II” and (c) “III”



(a) (b)

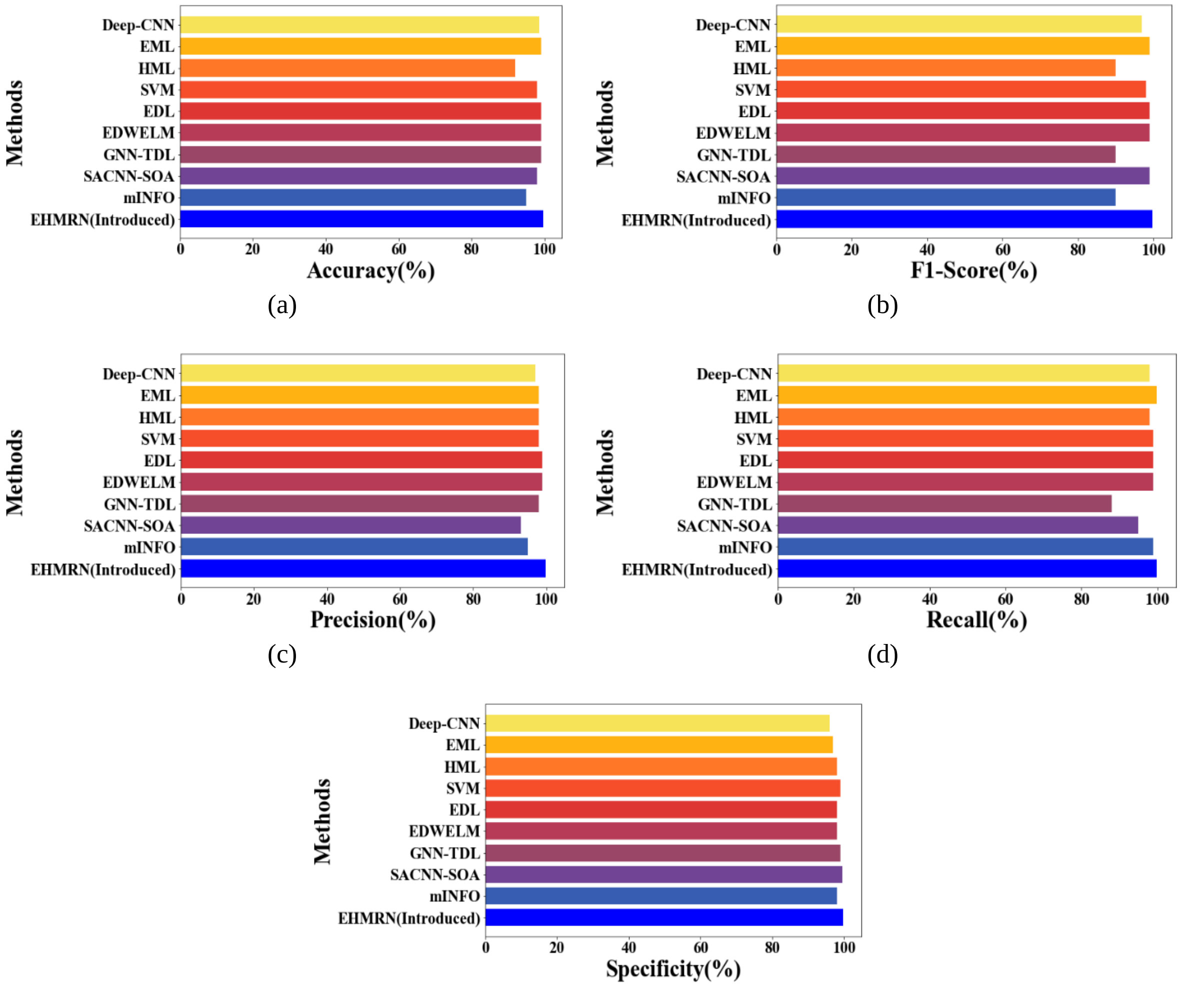
(c) (d)

(e) (f)

Figure 5. Training and validation curves of (a) accuracy using dataset "I", (b) loss using dataset "I", (c) accuracy using dataset "II", (d) loss using dataset "II", (e) accuracy using dataset "III", and (f) loss using dataset "III"

EHMRN exhibits superior accuracy compared to other optimization algorithms, as illustrated in Figure 9. This is attributed to the integration of MROA with hybrid network, preventing entrapment in local optima and facilitating the discovery of global optimum solutions, consequently leading to enhanced accuracy. EHMRN consistently achieves higher accuracy across all numbers of features, surpassing other algorithms by a significant margin. Figure 10 contrasts the execution time against the number of iterations for EHMRN and several traditional algorithms, including EML, SACNN-SOA and mINFO. Strikingly, despite being the most intricate and accurate algorithm among the considered methods, EHMRN exhibits the shortest execution time. This intriguing finding underscores not only the algorithm's complexity but also its efficiency, positioning EHMRN as a compelling choice for minimizing computational time while. The Table 2 presents Matthew's correlation coefficient (MCC) scores as percentages for various approaches, including Deep CNN, EDWELM and EHMRN. Deep CNN achieves a commendable 96% MCC, indicating robust predictive performance. EDWELM outperforms with an impressive 99.83% MCC, while EHMRN attains a perfect 100% MCC, highlighting its flawless predictive accuracy.

This implies that EHMRN excels in making precise predictions, showcasing its potential as a highly accurate classification model. Similarly, Table 3 shows Kappa scores for SVM, EDWELM and EHMRN. SVM achieves 98.67%, EDWELM surpasses with 99.66% and EHMRN attains a perfect 100%, indicating exceptional accuracy in classification tasks. Table 4 displays execution times for EML, SACNN-SOA, mINFO and EHMRN. EML records the shortest time at 0.119 seconds, indicating swift execution. SACNN-SOA and mINFO exhibit longer durations of 18 seconds and 8.0488 seconds, respectively. Notably, the introduced EHMRN method achieves the shortest execution time of 0.09 seconds, showcasing its efficiency and speed. The MROA incorporated into hybrid network contributes to its superior performance by ensuring better convergence speed compared to the other approaches.



(a) (b)

(c) (d)

(e)

Figure 6. Performance of (a) accuracy, (b) F1-score, (c) precision, (d) recall and (e) specificity comparison

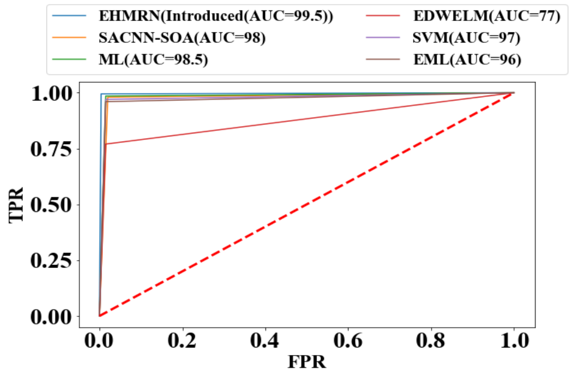


Figure 7. ROC curve

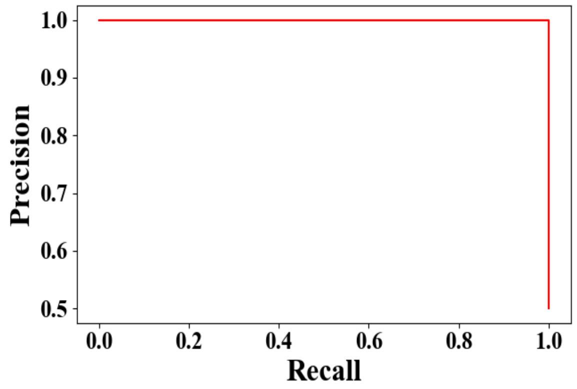


Figure 8. Precision-recall curve

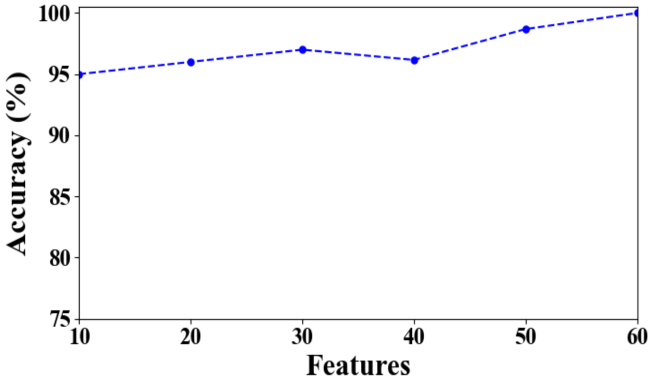


Figure 9. Accuracy vs features

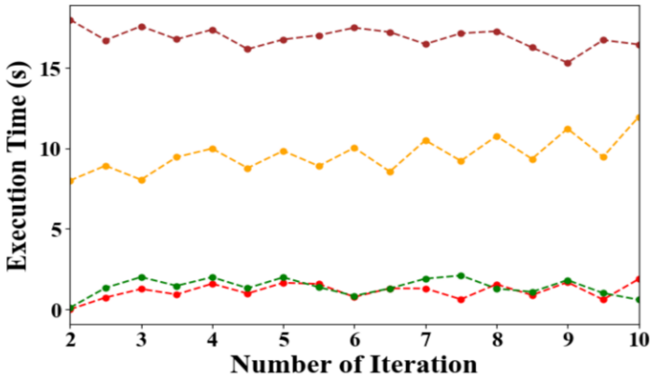


Figure 10. Execution time

Table 2. MCC comparison

Approaches	Deep CNN [11]	EDWELM [17]	EHMRN (Introduced)
MCC (%)	96	99.83	100

Table 3. Kappa comparison

Approaches	SVM [15]	EDWELM [17]	EHMRN (Introduced)
Kappa (%)	98.67	99.66	100

Table 4. Execution time comparison

Approaches	EML [12]	SACNN-SOA [21]	mINFO [20]	EHMRN (Introduced)
Time (s)	0.119	18	8.0488	0.09

4.1. Discussion

The existing problems in CKD detection include limited dataset sizes and the challenge of optimizing algorithms for enhanced accuracy and efficiency. The proposed method tackles these issues comprehensively. Firstly, the use of GAN for big data analytics expands the dataset size, addressing the problem of limited data availability, potentially enhancing model generalization and accuracy. Notably, the EHMRN, optimized via the MROA, demonstrated exceptional performance, achieving above 99.9% accuracy across key evaluation metrics by tuning the hyperparameter β and loss function L. This superiority was complemented by theoretical advancements, as EHMRN strategically avoided local optima pitfalls and efficiently navigated the solution space to identify global optimum solutions. Additionally, the application of MROA for tuning neural network parameters accelerates convergence and enhances accuracy, outperforming alternative methodologies. However, the comparative analysis highlights EHMRN's numerical superiority over established methods in CKD detection. Moreover, the incorporation of MEVO for feature selection underscored the theoretical soundness of the approach, ensuring the extraction of crucial features for optimizing CKD prediction models. Additionally, this research addressed privacy concerns by leveraging Spark for secure cloud-based storage, aligning with theoretical principles of ethical data handling in healthcare artificial intelligent research.

Enhanced hybrid mud ring network (EHMRN) consistently achieves exceptional scores across key evaluation metrics, surpassing Deep CNN, SVM, EDWELM, and other methods. Notably, EHMRN attains accuracy, precision, recall, and specificity scores above 99.9%, showcasing its outstanding predictive capabilities. Additionally, EHMRN achieves a perfect MCC score of 100%, indicating flawless predictive accuracy, while traditional methods like deep CNN achieve a commendable 96% MCC. EHMRN also outperforms in classification tasks, achieving a perfect Kappa score of 100%, while SVM and EDWELM achieve 98.67% and 99.66%, respectively. Moreover, EHMRN demonstrates superior execution times, with the shortest time of 0.09 sec compared to other algorithms, further emphasizing its efficiency and speed. These numerical comparisons underscore EHMRN's significant advancements in achieving highly accurate and efficient CKD detection results, establishing it as a benchmark for accurate and reliable prediction models in the field.

EHMRN consistently achieves exceptional scores across key evaluation metrics, surpassing Deep CNN, SVM, EDWELM, and other methods. Notably, EHMRN attains accuracy, precision, recall, and specificity scores above 99.9%, showcasing its outstanding predictive capabilities. Additionally, EHMRN achieves a perfect MCC score of 100%, indicating flawless predictive accuracy, while traditional methods like deep CNN achieve a commendable 96% MCC. EHMRN also outperforms in classification tasks, achieving a perfect Kappa score of 100%, while SVM and EDWELM achieve 98.67% and 99.66%, respectively. Moreover, EHMRN demonstrates superior execution times, with the shortest time of 0.09 seconds compared to other algorithms, further emphasizing its efficiency and speed. These numerical comparisons underscore EHMRN's significant advancements in achieving highly accurate and efficient CKD detection results, establishing it as a benchmark for accurate and reliable prediction models in the field. By combining numerical advancements with theoretical innovations, the research made a substantial contribution to the field of CKD detection, leading to the development of more accurate, efficient, and ethically sound healthcare AI solutions.

5. CONCLUSION

This research underscores the pivotal role of early-stage chronic kidney disease (ESDCKD) detection in facilitating timely intervention and enhancing understanding of disease progression. The introduction of EHMRN represents a significant breakthrough, effectively addressing existing limitations surrounding slow convergence, insufficient integration of advanced technologies, and privacy concerns. By leveraging extensive data collection through AI-based approaches and optimizing the hybrid network's performance with MROA, this study propels the prompt and accurate identification of CKDs in their nascent stages. Unlike previous methodologies, EHMRN effectively addresses several critical limitations, including slow convergence, inadequate integration of advanced technologies, and privacy concerns. By leveraging extensive data collection using AI-based approaches and optimizing the hybrid network's performance through MROA, this study pioneers a comprehensive and efficient approach to prompt and accurate CKD identification. Furthermore, the incorporation of Spark bolsters privacy measures, ensuring compliance with healthcare data regulations. The compelling results, boasting remarkable metrics including accuracy (99.96%), execution time (0.09 sec), MCC (100%), Kappa (100%), and ROC (1), underscore the remarkable advancement achieved in the realm of CKD detection. Moving forward, efforts to address interpretability challenges, enhance model robustness across diverse demographics, and validate methodologies through collaboration with healthcare institutions are imperative for realizing the real-world applicability and broader impact of ESDCKD detection techniques. Overall, this study heralds a new era in early CKD detection, promising improved patient outcomes and transformative contributions to healthcare delivery.

In future, addressing interpretability challenges, enhancing model robustness across demographics, and validating CKD detection methods through collaboration with healthcare institutions are crucial steps for their real-world application and broader impact. Overcoming interpretability hurdles ensures transparency in decision-making processes, while improving model robustness ensures equitable healthcare outcomes across diverse populations. Collaborations with healthcare institutions facilitate the integration of CKD detection tools into clinical practice, ultimately improving patient care and advancing public health efforts.

REFERENCES

[1] H. Zhang *et al.*, “Capability of intravoxel incoherent motion and diffusion tensor imaging to detect early kidney injury in type 2 diabetes,” *European Radiology*, vol. 32, no. 5, pp. 2988–2997, May 2022, doi: 10.1007/s00330-021-08415-6.

[2] R. J. Janse *et al.*, “Use of guideline-recommended medical therapy in patients with heart failure and chronic kidney disease: from physician’s prescriptions to patient’s dispensations, medication adherence and persistence,” *European Journal of Heart Failure*, vol. 24, no. 11, pp. 2185–2195, Nov. 2022, doi: 10.1002/ejhf.2620.

[3] M. Fontecha-Barriuso *et al.*, “Tubular mitochondrial dysfunction, oxidative stress, and progression of chronic kidney disease,” *Antioxidants*, vol. 11, no. 7, Jul. 2022, doi: 10.3390/antiox11071356.

[4] V. Arumugham, B. P. Sankaralingam, U. M. Jayachandran, K. V. S. S. R. Krishna, S. Sundarraj, and M. Mohammed, “An explainable deep learning model for prediction of early-stage chronic kidney disease,” *Computational Intelligence*, vol. 39, no. 6, pp. 1022–1038, Dec. 2023, doi: 10.1111/coin.12587.

[5] S. Lee *et al.*, “Machine learning–aided chronic kidney disease diagnosis based on ultrasound imaging integrated with computer-extracted measurable features,” *Journal of Digital Imaging*, vol. 35, no. 5, pp. 1091–1100, 2022, doi: 10.1007/s10278-022-00625-8.

[6] S. Chitra and V. Jayalakshmi, “Prediction of heart disease and chronic kidney disease based on internet of things using RNN algorithm,” in *Proceedings of Data Analytics and Management: ICDAM 2021*, 2022, pp. 467–479, doi: 10.1007/978-981-16-6289-8_40.

[7] A. Khade, A. V Vidhate, and D. Vidhate, “FFN-XGB-design of a hybrid feed forward neural network and extreme gradient boosting model for early prediction of chronic kidney disease,” *International Journal of System Assurance Engineering and Management*, Jun. 2023, doi: 10.1007/s13198-023-01993-2.

[8] X. Yan *et al.*, “Establishment and evaluation of artificial intelligence-based prediction models for chronic kidney disease under the background of big data,” *Evidence-Based Complementary and Alternative Medicine*, vol. 2022, pp. 1–7, Jul. 2022, doi: 10.1155/2022/6561721.

[9] S. Patil and S. Choudhary, “Hybrid classification framework for chronic kidney disease prediction model,” *The Imaging Science Journal*, vol. 72, no. 3, pp. 367–381, Apr. 2024, doi: 10.1080/13682199.2023.2206272.

[10] J. M. Thomas, B. M. Huuskes, C. G. Sobey, G. R. Drummond, and A. Vinh, “The IL-18/IL-18R1 signalling axis: diagnostic and therapeutic potential in hypertension and chronic kidney disease,” *Pharmacology & Therapeutics*, vol. 239, Nov. 2022, doi: 10.1016/j.pharmthera.2022.108191.

[11] S. Rajeashwari and K. Arunesh, “Chronic disease prediction with deep convolution based modified extreme-random forest classifier,” *Biomedical Signal Processing and Control*, vol. 87, Jan. 2024, doi: 10.1016/j.bspc.2023.105425.

[12] M. Mustafizur Rahman, M. Al-Amin, and J. Hossain, “Machine learning models for chronic kidney disease diagnosis and prediction,” *Biomedical Signal Processing and Control*, vol. 87, Jan. 2024, doi: 10.1016/j.bspc.2023.105368.

[13] X.-Y. Ge, Z.-K. Lan, Q.-Q. Lan, H.-S. Lin, G.-D. Wang, and J. Chen, “Diagnostic accuracy of ultrasound-based multimodal radiomics modeling for fibrosis detection in chronic kidney disease,” *European Radiology*, vol. 33, no. 4, pp. 2386–2398, Dec. 2022, doi: 10.1007/s00330-022-09268-3.

[14] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, “Chronic kidney disease prediction based on machine learning algorithms,” *Journal of Pathology Informatics*, vol. 14, 2023, doi: 10.1016/j.jpi.2023.100189.

[15] D. Swain *et al.*, “A robust chronic kidney disease classifier using machine learning,” *Electronics*, vol. 12, no. 1, Jan. 2023, doi: 10.3390/electronics12010212.

[16] D. M. Alsekait *et al.*, “Toward comprehensive chronic kidney disease prediction based on ensemble deep learning models,” *Applied Sciences*, vol. 13, no. 6, Mar. 2023, doi: 10.3390/app13063937.

[17] T. B. Prasad Reddy and D. Vydeki, “Ebola deep wavelet extreme learning machine based chronic kidney disease prediction on the internet of medical things platform,” *Concurrency and Computation: Practice and Experience*, vol. 35, no. 1, Jan. 2023, doi: 10.1002/cpe.7446.

[18] P. K. Rao, S. Chatterjee, K. Nagaraju, S. B. Khan, A. Almusharraf, and A. I. Alharbi, “Fusion of graph and tabular deep learning models for predicting chronic kidney disease,” *Diagnostics*, vol. 13, no. 12, Jun. 2023, doi: 10.3390/diagnostics13121981.

[19] S. Pal, “Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms,” *Multimedia Tools and Applications*, vol. 82, no. 26, pp. 41253–41266, Nov. 2023, doi: 10.1007/s11042-023-15188-1.

[20] E. H. Houssein and A. Sayed, “A modified weighted mean of vectors optimizer for chronic kidney disease classification,” *Computers in Biology and Medicine*, vol. 155, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106691.

[21] J. Sulthan Alikhan, R. Alageswaran, and S. Miruna Joe Amali, “Self-attention convolutional neural network optimized with season optimization algorithm espoused chronic kidney diseases diagnosis in big data system,” *Biomedical Signal Processing and Control*, vol. 85, Aug. 2023, doi: 10.1016/j.bspc.2023.105011.

[22] L. Rubini, P. Soundarapandian, and P. Eswaran, “Chronic kidney disease,” *UCI Machine Learning Repository*, 2015, doi: 10.24432/C5G020

[23] A. K. Pandit, “Chronic kidney disease,” *Kaggle*, https://www.kaggle.com/datasets/abhia1999/chronic-kidney-disease (accessed Jan. 03, 2023).

[24] C. Kathuria, “Chronic kidney disease - explored !” *Kaggle*, https://www.kaggle.com/code/chayan8/chronic-kidney-disease-explored (accessed Jan. 03, 2023).

[25] V. A. Fajardo *et al.*, “On oversampling imbalanced data with deep conditional generative models,” *Expert Systems with Applications*, vol. 169, May 2021, doi: 10.1016/j.eswa.2020.114463.

[26] C. Wongoutong, “Imputation methods for missing response values in the three parts of a central composite design with two factors,” *Journal of Statistical Computation and Simulation*, vol. 92, no. 11, pp. 2273–2289, 2022, doi: 10.1080/00949655.2022.2027424.

[27] K. A. Tarnowska, A. Bagavathi, and Z. W. Ras, “High-performance actionable knowledge miner for boosting business revenue,” *Applied Sciences*, vol. 12, no. 23, Dec. 2022, doi: 10.3390/app122312393.

[28] M. A. Azad *et al.*, “Energy valley optimizer (EVO) for tracking the global maximum power point in a solar PV system under shading,” *Processes*, vol. 11, no. 10, Oct. 2023, doi: 10.3390/pr11102986.

[29] H. Nasiri *et al.*, “Classification of covid-19 in chest X-ray images using fusion of deep features and LightGBM,” in *2022 IEEE World AI IoT Congress (AIIoT)*, Jun. 2022, pp. 201–206, doi: 10.1109/AIIoT54504.2022.9817375.

[30] E. A.-R. Hamed, M. A.-M. Salem, N. L. Badr, and M. F. Tolba, “An efficient combination of convolutional neural network and LightGBM algorithm for lung cancer histopathology classification,” *Diagnostics*, vol. 13, no. 15, Jul. 2023, doi: 10.3390/diagnostics13152469.

[31] A. S. Desuky, M. A. Cifci, S. Kausar, S. Hussain, and L. M. El Bakrawy, “Mud ring algorithm: a new meta-heuristic optimization algorithm for solving mathematical and engineering challenges,” *IEEE Access*, vol. 10, pp. 50448–50466, 2022, doi: 10.1109/ACCESS.2022.3173401.

BIOGRAPHIES OF AUTHORS

Mamatha B. is working as an assistant professor in the CSM (AI and ML), CMR Technical Campus Medchal, Telangana. She received her B.Tech. degree in 2012 and M.Tech. degree in 2014 from Visveswaraya Technological University, Belgaum, India. Currently, she is pursuing her research at PDA College of Engineering, Gulbarga, Karnataka, India. Her fields of interest are bigdata, machine learning, and deep learning. She can be contacted at email: mamatha.789@gmail.com.

Sujatha P. Terdal is working as professor and HOD in Department of Computer Science and Engineering, P.D.A College of Engineering Gulbarga. She has completed B.E. from Gulbarga University, Gulbarga. She received her M.Tech. degree from Visveswaraya Technological University, Belgaum, India, in 2002. She has completed PhD degree from JNTU, Hyderabad in 2014. She has attended several national and international conferences/seminars and presented papers. She has organized many workshops, short term courses and other refresher and orientation programs. Her research interests include network security, mobile ad hoc networks, wireless sensor networks, cloud computing, machine learning. She can be contacted at email: sujathapterdal@outlook.com.