Journal of Intelligent Systems and Internet of Things
Vol. 15, No. 01, PP. 122-132, 2025
DOI: https://doi.org/10.54216/JISIoT.150110

Optimized KNN Algorithm for Diabetic Retinopathy Classification with PCA-Based Data Fusion and Cuckoo Search Optimization

Ali Azawii Abdul Lateef [1,*], Ahmed Subhi Abdalkafor[2], Ahmed Adil Nafea[3]

[1]University Headquarter, Department of Human Resources, University Of Anbar, Anbar, Iraq

[2]College of Computer Science and Information Technology, University Of Anbar, Anbar, Iraq

[3]Department of Artificial Intelligence, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq

Emails: aliazawii@uoanbar.edu.iq; ahmed.abdalkafor@uoanbar.edu.iq; ahmed.a.n@uoanbar.edu.iq

Abstract

Diabetes is a disease that occurs when the body is unable to use the insulin it produces effectively or the body fails to produce enough insulin. One of the most important complications of this disease is diabetic retinopathy (DR), which is considered the main cause of severe visual impairment and blindness. Previous studies have proven that the KNN algorithm is an effective algorithm for solving classification and prediction problems, as the performance of this algorithm rely on determining the value of the K parameter because the inappropriate choice of this value can negatively affect the accuracy of classification. On the other hand, adjusting this value manually is very difficult because this value depends on the state of determining the solution to the problem each time. Therefore, there is still an urgent need to use smart algorithms to adjust this value and obtain an ideal value that ultimately leads to obtaining a very high classification accuracy. In this paper, the Cuckoo Search algorithm was used, which is considered one of the smart and modern algorithms in the field of diagnosis, in addition to applying more than one technique and algorithm to build an integrated system to enhance the accuracy of diagnosis and obtain competitive diagnostic accuracy. The proposed work was implemented using the Debrecen diabetic retinopathy dataset and competitive results were obtained for recall, sensitivity, precision, F1 score, accuracy and specificity (98.05%), (97.30%), (99.01%), (98.70%), (99.70%), and (99.08%), respectively. Our results demonstrate that the Cuckoo Search algorithm is an effective and suitable choice for optimizing the parameters in the KNN algorithm, in addition to enhancing this algorithm to diagnose the disease early and support direct intervention and treatment, and this method lays the foundation for diagnosing other diseases and thus improving patient care in most related fields.

Received: July 02, 2024 Revised: September 27, 2024 Accepted: December 25, 2024

Keywords: Diabetes Retinopathy; Data Fusion; PCA-transformed features; Cuckoo Search Optimization; Optimized KNN algorithm

1. Introduction

Diabetic retinopathy (DR) is a microvascular problem of diabetes that can lead to severe vision impairment and blindness[1]. It is primarily caused via uncontrolled hyperglycaemia, which induces microangiopathy and results in destruction to the retinal blood vessels. DR can be classified into two distinct stages the first one is non-proliferative and the other is proliferative DR. One of the most relating indications of this disease is macular edema, which often presents with no early symptoms and can lead to sudden vision loss[2]. The prevalence of DR has get higher much in recent years, with projections implying that the number of affected individuals will increase from 171 million in 2000 to an estimated 366 million by 2030[3]. Additionally, it was reported that in 2019, DR accounted for the deaths of 1.5 million individuals, highlighting its status as an important public health issue [4],[5]. Artificial intelligence has entered many fields such as medicine [6], education [7], industry [8], agriculture

[9], and others, due to its ability to process data quickly through either prediction, classification, or discovery. This has led researchers to prepare and create datasets in various fields [10][11]to facilitate these processes.

Given the alarming rise in DR cases, there has been a concerted effort to discover the applications of artificial intelligence (AI) in its diagnosis. DR is well-suited for AI interventions due to the availability of large datasets, early successes in AI applications for diabetes, and AI systems achieving accuracy comparable to or surpassing that of human specialists. Focusing on DR highlights AI's broader potential to improve diagnostics for various medical conditions [12].

Recent advancements in imaging technologies and machine learning have opened new chances for enhancing the accuracy of DR classification. Although these developments, challenges persist in achieving high classification accuracy due to the complexities of the data involved. Large and high-dimensional feature sets can obscure relevant information, making it difficult for models to generalize well across different datasets. Also, lasting dimensionality reduction techniques often sacrifice important data variance, adversely impacting the performance of classification algorithms.

To address these issues, this paper proposed an integrated classification model that employs dimensionality reduction techniques, including PCA, to maintain significant variance while reducing complexity. Additionally, we utilize MIFS to identify the most relevant features that contribute to classification accuracy. Finally, we optimize the parameters of the KNN algorithm using the Cuckoo algorithm, aiming to achieve competitive accuracy levels. The outcomes of this study not only advance the current methodologies for DR classification but also pave the way for improving diagnostic accuracy in medical imaging tasks more broadly.

This paper is organized as follows: The introduction outlines the significance of accurate classification in DR and the motivation behind our research. The related work section reviews classification techniques and their limitations. The methodology section details proposed model, including PCA, MIFS, and the Cuckoo algorithm for optimizing KNN. The experimental setup describes the datasets and metrics utilized for evaluation. In the results and discussion section, shows findings and analyze implications. Finally, the conclusion summarizes proposed contributions and future research directions.

2. Related Work

Yihao Li et al. [13] used of multimodal data fusion for retinal disease classification, focusing on glaucoma and diabetic retinopathy. The authors use three deep learning-based fusion strategies: early fusion, intermediate fusion, and hierarchical fusion. They argue that complementary information between media is not fully exploited in early and intermediate mergers. Therefore, they developed an approach that explores the correlations between media and combines features across multiple dimensions called the hierarchical approach. This method is applied to both public and private datasets, achieving good performance.

Yihao et al. [14] this work aims to improve the screening and diagnosis of DR, a leading cause of blindness in developed countries. The "Evaluation intelligent de la rétinopathie diabétique" (EviRed) project used AI to enhance the outdated classification system. This study utilizes modern fundus imaging devices and patient medical data to assist ophthalmologists in making more accurate diagnoses and predictions during DR follow-up. The fusion of multiple imaging modalities, including 3D structural optical coherence tomography, 3D OCT angiography, and 2D Line Scanning Ophthalmoscope, is employed for automatic detection of proliferative DR.
Daho et al. [15] this paper presented a novel multimodal deep learning method for enhancing DR diagnosis, a severe complication of diabetes. Using the fusion of 3D-ResNet50 and ResNet50 models, high-resolution 3D OCTA data is fused with UWF-CFP images. The authors also incorporate a multimodal extension of the Manifold Mixup technique to improve model generalization. Experimental results show that the proposed approach significantly improves DR classification performance compared to single-modality methods. This work contributes to the growing body of research demonstrating the potential of multimodal fusion for improving early detection and diagnosis of DR, ultimately leading to better clinical outcomes.

Nneji et al. [16] this paper aims to classify DR stages in adults aged 25-74, a leading cause of visual impairment. Early detection is important to prevent proliferative DR, which can lead to severe vision loss. The authors proposed a Weighted Fusion Deep Learning Network (WFDLN) to enhance DR stage classification. The framework processes two channels of fundus images, fused utilizing a weighted approach, and soft-max classification to determine DR stages. The WFDLN model achieved good accuracy levels, comparable to other state-of-the-art models, contributing to the development of reliable DR stage classification systems.

Using three diabetes datasets, Bekim et al. [17] investigated the performance of three algorithms Naïve Bayes (NB), J48, ML and Random Forest (RF), counting two from the public domain and one from a research study. This work highlights the potential of ML to analyze vast amounts of medical data, particularly in the context of the global diabetes epidemic, where current predictive performance is inadequate. The authors address the challenges faced via ML in healthcare and contribute to bridging gaps in existing research on diabetes prediction.

Sisodia et al.[18] focus on developing a ML model to accurately predict diabetes, a chronic disease with serious complications if untreated. The study uses three classification algorithms (NB, SVM, and DT) applied to a diabetic patient's database from the UCI repository. The system was implemented and tested using accuracy measures, where the NB classifier reached the highest accuracy of 76.30%., underscoring the potential of ML in early diabetes detection.

While recent studies have significantly advanced the classification and detection of DR through various multimodal approaches, there remains a critical gap in achieving high accuracy across diverse datasets. Thus, this paper aims to bridge this gap by proposing an integrated classification model that combines dimensionality reduction techniques with an optimized KNN algorithm, striving to enhance detection accuracy in DR classification, thereby contributing to more reliable and clinically relevant outcomes in ophthalmology.

3. Methodology

The proposed work steps that were implemented in building this system to obtain competitive accuracy are summarized in Figure (1). These proposed steps are explained and discussed in the following subsections of this paper.

3.1 Prepossessing

This step is very important in the dataset, including cleaning, converting, formatting and organizing it into a format that makes it suitable for training and testing, with the aim of improving the quality of this data and the model's ability to learn efficiently

3.1.1 Principal Component Analysis (PCA)

Principal component analysis is a practical mathematical technique that has been used in many different domains. It assists in minimizing the data set's dimensionality and using graphical representation to find patterns in the data. Principal component analysis is a data reduction approach, to put it briefly. PCA aims to identify a restricted number of linear groups (principal components) of a set of variables while conserving as much information as possible in the original variables.

Plotting, regression, clustering, and other techniques may frequently be performed using a limited set of main components rather than the original variables. Another way to think of principal component analysis is as a method for eliminating multicollinearity from data. This method converts the initial collection of variables into a new set of principle components, which are uncorrelated random variables in order for the first principal component to give an explanation of the dissimilarity in the primary data as possible, these new variables are linear groups of the original variables and are drawn in descending order of significance. Although this technique is most useful in high dimensional datasets, PCA is widely employed across all scientific fields. This method allows for the transformation of a dataset into a new coordinate system in which the greatest variance occurs within the first few coordinates. Overall, PCA serves various purposes, including simplifying data while retaining its essential features, recognizing patterns, and compressing the original dataset. To illustrate the mathematical operation of this technique, as in the equations below [19]:

$$Cov_{Mat} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} (\mathcal{X}_i - \bar{\mathcal{X}})(\mathcal{X}_i - \bar{\mathcal{X}})^{\mathcal{T}} \quad (1)$$

where the $\mathcal{X}_i$, $\bar{\mathcal{X}}$ are the data points and mean vector respectively.

Decomposition of Eigenvalue:

Calculate the eigenvectors $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3, \dots\dots\dots\dots\dots, \mathcal{V}_d$ and corresponding eigenvalues $\lambda_1, \lambda_2, \lambda_3, \dots\dots\dots\dots\dots\dots, \lambda_d$ of Cov_{Mat}.

The projection matrix $P\_M_{\mathcal{K}}$ is formed by assigning the top eigenvectors $\mathcal{K}$ that correspond to the largest eigenvalues

$$P\_M_{\mathcal{K}} = [\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3, \dots\dots\dots\dots\dots, \mathcal{V}_{\mathcal{K}} \quad (2)$$

The data transformation D_T_{PCA} is obtained through:

$$D\_T_{PCA} = D\_T * P\_M_{\mathcal{K}} \quad (3)$$

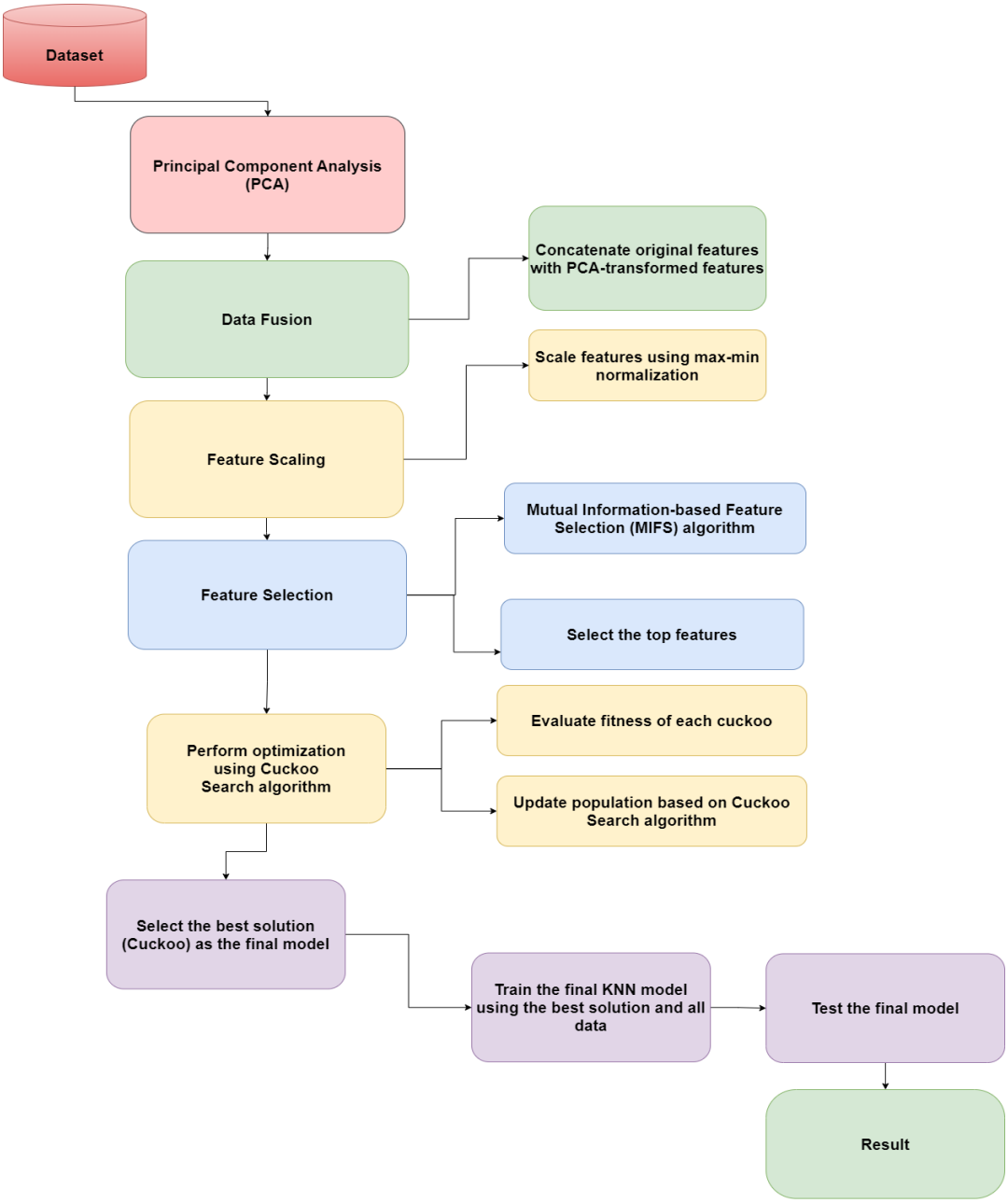

Figure 1. Block diagram of Proposed System

3.2. Data Fusion

The proposed model can access both the transformed data and the raw data that represent the main data patterns by combining the transformed features and the original features using PCA [20]. This method is used in our proposed work because it is useful for the algorithm applied in the subsequent steps, as these algorithms learn better using rich features.

3.3. Feature Scaling

This process is one of the important processes in measuring features to unify the range of features in the dataset, as this method works so that each feature contributes in a balanced way to training the proposed model and deletes the features that have large ranges from controlling the features with small ranges[21], as this process was applied because the next step in our proposed work is to apply the KNN algorithm, which depends on the distance and whose goal or focus is to weigh the features with the largest range significantly.

3.3.1. Max-Min Normalization

This technique works to standardize the data so that the range of values in the dataset is converted to a new range. This technique is very useful for datasets that contain widely varying features. This technology aims to make parallel or equal scales in order to facilitate processing within the system. Below is the mathematical operation of this technique[22].

$$\mathcal{X}\text{norm} = \frac{\mathcal{X}-\mathcal{X}_{min}}{\mathcal{X}_{max}-\mathcal{X}_{min}} \quad (4)$$

where $\mathcal{X}_{max}$ and $\mathcal{X}_{max}$ are maximum and minimum values of $\mathcal{X}$, respectively.

3.4 Feature Selection

This process focuses on identifying the most relevant features in the database that significantly enhance the performance of the model. This step was used in our work because choosing the right features leads to better accuracy, which is what we achieved, as irrelevant or redundant features can introduce noise and thus lead to lower performance or degrade the system completely. In addition, this step works to reduce overfitting because fewer features lead to the possibility of the model learning less from noise. Finally, this process reduces training time and enhances the model's ability to understand because choosing important and relevant features is easier to interpret and understand, especially in applications related to health matters[23].

3.5.1 Mutual Information-Based Feature Selection (MIFS)

This technique is designed to improve the performance of models by reducing the number of features and only keeping the most related or relevant features, as this technique depends on mutual information by measuring the amount of information and dependency between features. This technique was chosen in our proposed work because it is characterized by reducing useless features, which ultimately leads to obtaining two basic advantages, namely improving the proposed work performance and the speed of its implementation, in addition to reducing the computational complexity of this system. Below are the mathematical operations for this technique [24].

Determine the MI among the each feature $\mathcal{F}_i$ the target $\mathcal{Y}$.

$$M\_I(\mathcal{F}_i;\mathcal{Y}) = \sum_{\mathcal{f}\in\mathcal{F},\mathcal{y}\in\mathcal{Y}} \mathcal{P}(\mathcal{f}_i,\mathcal{y}) \log(\frac{\mathcal{P}(\mathcal{f}_i,\mathcal{y})}{\mathcal{P}(\mathcal{f}_i)\,\mathcal{P}(\mathcal{y})} \quad (5)$$

where $\mathcal{P}(\mathcal{f}_i,\mathcal{y})$ and $\mathcal{P}(\mathcal{f}_i)\,\mathcal{P}(\mathcal{y})$ are the distribution of joint probability and probabilities of marginal respectively.
MRMR Criterion: By minimizing redundancy, MRMR seeks to maximize the MI between the target and the selected features.

$$\text{MRMR}(\mathcal{S}) = \frac{1}{\mathcal{K}}\sum_{i=1}^{\mathcal{K}}\mathcal{I}(\mathcal{F}_{\mathcal{S}};\mathcal{Y}) - \frac{1}{\mathcal{K}^2}\sum_{i=1}^{\mathcal{K}}\sum_{j=1}^{\mathcal{K}}\mathcal{I}(\mathcal{F}_{\mathcal{S}_i};\mathcal{F}_{\mathcal{S}_j}) \quad (6)$$

where $\mathcal{S}$ and $\mathcal{K}$ are the subset of selected feature the number of selected features respectively, while $\mathcal{F}_{\mathcal{S}_i};\mathcal{F}_{\mathcal{S}_j}$ are features in subset $\mathcal{S}$.

3.5 Cuckoo Search Algorithm

This algorithm is based on a method inspired by the behavior of birds in laying eggs in the nests of other birds. This algorithm is built on three basic processes. The first is that each egg is considered a potential solution within the scope of the problem to be improved, so that each egg represents a set of values that the algorithm seeks to reach. The second process focused on evaluating the quality of solutions by evaluating each solution using the objective function, and the solutions that achieve the highest value for this function are kept. The last process, which is considered one of the important processes, focused on replacing the worst solutions through a Lévy Flight, which aims to explore the solution space more deeply and broadly, which increases the chances of finding the optimal solution among these solutions, this algorithm was chosen in our work because it is easy to implement, as it does not require very large modifications of the parameters. Below are the mathematical operations for this algorithm [25].
Lévy Flight: Apply Lévy flights to create new solutions (number of neighbors):

$$\mathcal{X}_{t+1} = \mathcal{X}_t + \alpha.\mathcal{L}_t.\mathcal{U} \quad (7)$$

3.6 K-Nearest Neighbours (KNN)

This algorithm depends on classifying new samples based on the similarity between these samples and other samples in the data on which the model was trained, based on the distance between those samples to determine the degree of similarity. This algorithm depends on four basic operations: first, calculating the distance between the new sample and the rest of the samples, then determining the closest K to the new sample, then determining the most frequently repeated sample, and finally generating the result based on the neighbors' votes. Below is an explanation of the mathematical operation of this algorithm[26].
Distance Calculation: Calculate the distances to each training instance (PPP) for a given query instance (QQQ).

$$Dec(\mathcal{Q}, \mathcal{P}_i) = \sqrt{\sum_{j=1}^{n}(\mathcal{Q}_j - \mathcal{P}_{ij})^2} \quad (8)$$

where $\mathcal{Q}_j$, $\mathcal{P}_{ij}$ are features of $\mathcal{Q}$ and $\mathcal{P}_i$ respectively.
Majority Voting: Assign a class $\mathcal{Q}$ according to its $\mathcal{K}$ closest neighbors' majority class.

where $\boldsymbol{\mathcal{X}_t}$ and $\boldsymbol{\mathcal{X}_{t+1}}$ are the current and new solutions, $\boldsymbol{\alpha}$ and $\boldsymbol{\mathcal{L}_t}$ are scaling factor and Lévy flight step, while $\boldsymbol{\mathcal{U}}$ is a random vector from distribution of a Lévy.

Optimization of KNN algorithm

In this step, the cuckoo algorithm is implemented to find the optimal number of neighbours for the KNN algorithm, where Lévy flights are used to find new solutions. This process (the optimization process) updates the set of solutions more than once depending on the suitability as in the equation below.
PCA: By eigen decomposing the covariance matrix, the major components are found.

$$Cov_{Mat} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}}(\mathcal{X}_i - \bar{\mathcal{X}})(\mathcal{X}_i - \bar{\mathcal{X}})^{\mathcal{T}} \quad (9)$$

where the $\mathcal{X}_i$, $\bar{\mathcal{X}}$ are the data points and mean vector respectively.

Population Initialization: Randomly initialize (n) cuckoos with random solutions (number of neighbors) $\mathcal{K}_i$
Objective Function: Use KNN to evaluate each cuckoo's fitness (classification accuracy):

$$\text{fitness } (\mathcal{K}_i) = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Sorting: To prioritize better solutions, sort the cuckoos in descending order of fitness.
Lévy Flights: Create a new solution by applying Lévy flights to investigate the search space:

$$step_{size} = \alpha.levy\_flight() \quad (11)$$

$$new\_sol = step_{size} + \mathcal{K}_i \quad (12)$$

Where α is the scaling factor of step size and $levy\_flight()$ is the features two random variables, u and v, that are sampled from a normal distribution.
Update Population: Using new solutions by $levy\_flight()$ that replaced the worst solutions ensured that the solutions remained within the specified limits.
KNN Model Training and Evaluation: Used the best result from Cuckoo Search algorithm to train KNN.

$$Final\ Model\ of\ KNN = Fit\_c\_KNN(\mathcal{X}_{selected}, Y, \text{Num}_{\text{neighbors}}, \mathcal{K})$$

Where $\mathcal{K}$ is the optimal number of neighbors as determined via Cuckoo Search algorithm.
Prediction and Accuracy:
Using the trained KNN model, make predictions and determine accuracy.

$$ACC = \frac{Correct\ predict}{total\ predict} * 100\% \quad (13)$$

4. Results and Discussion

4.1 Dataset

The UCI Machine Learning Repository contains a diabetic retinopathy dataset available for testing purposes consisting of twenty features [27]. The process of the appearance or non-appearance of retinopathy is done by taking the features from the images in the data set by what these features provide of the presence of an injury or not. In the data set, the features are numbered from zero to nineteen. The quality of the image is measured by

feature number 0, which is represented by two values, either 0 or 1. The value 0 means the image is of poor quality and vice versa. As for feature number 1, it represents a pre-screening for the disease. Microtomy (MA) is represented by features from 2 to 7. In addition, the values associated with secretions are determined by features from 8 to 15. Information about the Euclidean distance and the optic disc is determined by feature number 16, and feature number 17 is responsible for the diameter of the optic disc. As for the modifications to AM and FM, they are binary values determined by feature number 18. Finally, feature 19 is a binary value, value number 1 represents the presence of symptoms and vice versa.

Performance Metrics: To calculate metrics including precision, recall, F1-score, sensitivity, and specificity, compute the confusion matrix [28].

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

$$\mathrm{F1-score} = \frac{2*\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \quad (16)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (17)$$

$$Specificity = \frac{TN}{TN+FP} \quad (18)$$

where FN, TN, FP and TP are the, False Negatives, True Negatives, False Positives, and True Positives respectively. The proposed model, which dimensionality reduction techniques, feature selection, and optimization algorithms, demonstrates substantial performance in classifying DR. As shown in Table 1 the final model achieved a high accuracy of 98.05%, indicating a highly effective ability in distinguishing among different stages of DR.

Table 1: Proposed Model Results

	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
Proposed Model	98.05%	97.30	99.01	98.70	99.70	99.08

In addition to accuracy, various other performance metrics were evaluated to provide a comprehensive assessment of the model's efficiency as shown in Figure (2).

The model achieved a precision of 97.30%, suggesting that the proportion of true positive predictions among all positive predictions is high, thereby minimizing false positives in the classification of DR. A remarkable recall of 99.01% shows the ability of model to correctly identify nearly all cases of DR, highlighting its efficiency in detecting the condition. With an F1 score of 98.70%, the model balances precision and recall effectively, showcasing its strength in classification tasks. The sensitivity level reached 99.70%, highlighting the model's ability to identify patients with DR correctly, which is important for early intervention and treatment. Additionally, the specificity of 99.08% reflects the model’s proficiency in accurately classifying patients.

The integration of PCA was instrumental in preserving the significant variance of the dataset while simplifying the feature set, allowing the model to focus on the most informative characteristics of the data. Furthermore, the implementation of MIFS proved effective in pinpointing the most relevant features that directly contribute to improving classification accuracy. By filtering out irrelevant or redundant features, MIFS helped enhance the model's overall performance and interpretability .

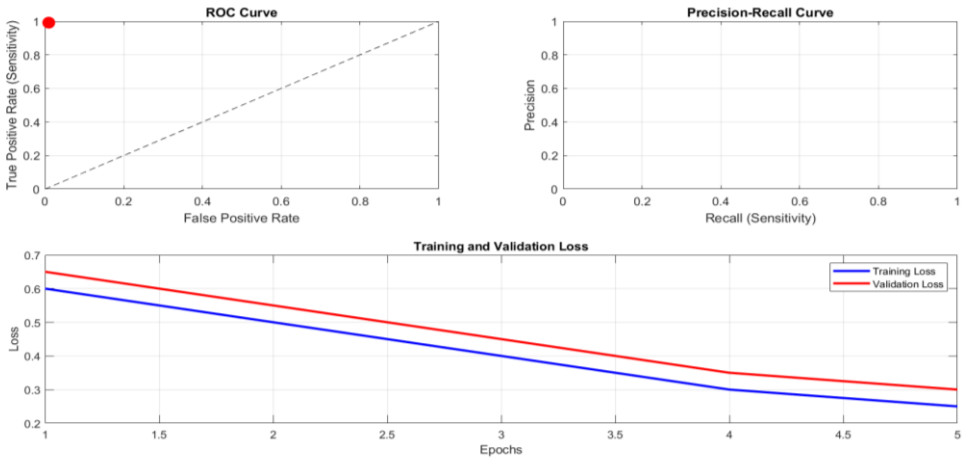

Figure 2. Model Performance Metrics

The proposed model also achieved a very strong successful performance, as proven by the AUC ratio, which reached 0.99140, which proves that the model is able to distinguish between positive and negative classes with a high degree of accuracy across different threshold settings. This value, which was close to the perfect degree, reflects that the model has a very high ability to correctly classify both true positives and negatives, and vice versa in terms of false negatives and positives.

The optimization of the KNN algorithm through the Cuckoo algorithm also played an important role in achieving these competitive accuracy levels. The Cuckoo algorithm's ability to intelligently explore the parameter space facilitated the identification of optimal hyper-parameters, leading to better decision boundaries and improved classification outcomes. Overall, the results display that the proposed model significantly advances current methodologies for DR classification, offering a capable approach to improve diagnostic accuracy in medical imaging. This performance not only enhances the potential for early detection and intervention in diabetic retinopathy but also sets the groundwork for future applications of similar methodologies in other medical conditions, ultimately contributing to better patient care and outcomes in ophthalmology and beyond.

The confusion matrix (CM) offers a clear illustration of the integrated classification performance of model in detecting DR. Figure 3 shows that the model correctly identified 246 TP, indicating a strong capability in detecting positive cases of DR. The number of FP is notably low at 7, reflecting a minimal rate of misclassification of negative cases as positive. Additionally, the model achieved a high number of TN, correctly classifying 735 instances as not having DR, which highlights its effectiveness in ruling out the disease when it is absent. However, the model did report 2 FN, where positive cases were incorrectly classified as negative, though this number remains minimal. Overall, the confusion matrix illustrates that the model has a high accuracy level, effectively distinguishing between the presence and absence of DR while minimizing classification errors.

Training Set			
TARGET / OUTPUT	Class0	Class1	SUM
Class0	246 24.85%	7 0.71%	253 97.23% 2.77%
Class1	2 0.20%	735 74.24%	737 99.73% 0.27%
SUM	248 99.19% 0.81%	742 99.06% 0.94%	981 / 990 99.09% 0.91%

Figure 3. Confusion Matrix

In this study, our proposed application achieves an impressive classification accuracy of 98.05% through the integration of dimensionality reduction techniques and optimization algorithms. By employing PCA to retain the most significant variance in the feature set, coupled with MIFS to identify the most relevant features, we have created a robust model. Furthermore, optimizing the KNN algorithm using the Cuckoo search algorithm significantly enhances performance. When comparing our results to state-of-the-art studies as shown in Figure 4, we find that our model outperforms several notable approaches.

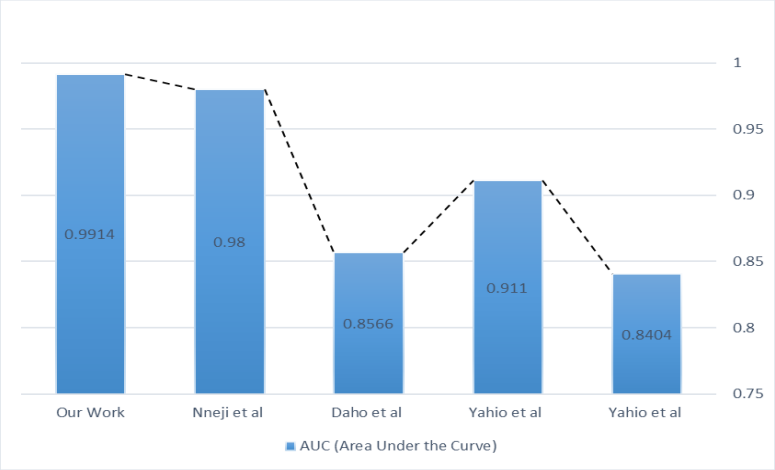

Figure 4. Area under the Curve

For instance, Yahio et al. [13] achieved an AUC of 0.8404 using a ResNet50 model with hierarchical fusion, while Yahio et al. [14] reached an AUC of 0.911 with Densenet121. Daho et al. [15] reported an AUC of 0.8566 using a deep multimodal fusion model for diabetic retinopathy (DR) severity evaluation. Nneji et al. [16], utilizing a weighted fusion deep learning network, achieved an AUC of 0.98, which is noteworthy but still falls short of our overall accuracy metric. Additionally, Fetaji et al. [18] demonstrated that among three machine learning algorithms (DT, SVM, and NB), NB shows the highest accuracy of 76.30%, which is considerably lower than our findings.

The superior performance of our model suggests that the combination of PCA and MIFS for feature selection, alongside the Cuckoo optimization of KNN, effectively addresses the challenges associated with high-dimensional data. This positions our approach as a promising alternative in the quest for enhanced classification accuracy in similar domains, highlighting the potential of integrating traditional machine learning techniques with advanced optimization strategies.

5. Conclusion

Failure to diagnose retinopathy early causes blindness and vision loss, which is a very important matter. Min-Max normalization was applied to measure the set of original features with the transformed by the PCA in order to ensure that these features range between zero and one. The most important features were identified on the measured data using the MIFS technique and the KNN algorithm was improved after applying the Cuckoo search algorithm, which is the most important and powerful part of this paper, where the suitability is evaluated based on the accuracy of the trained K-NN algorithm models. The best solution was used to train the KNN algorithm on the selected features and obtain a competitive diagnosis accuracy reached 98.05%. It is concluded from this proposed work on the importance of detecting and treating this disease early, and this work demonstrates the power of smart algorithms in diagnosing diseases. As a future work and as a result of the results obtained, there is a possibility of applying other machine learning algorithms and implementing the Cuckoo search algorithm to improve the structure of these algorithms and obtain a high competitive accuracy and apply it to diagnose other diseases.

Funding: None.

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] M. Kropp et al., "Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation," *EPMA J.*, vol. 14, no. 1, pp. 21–42, 2023. DOI: https://doi.org/10.1007/s13167-023-00314-8.

[2] D. S. Pushparani, J. Varalakshmi, K. Roobini, P. Hamshapriya, and A. Livitha, "Diabetic Retinopathy—A Review," *Curr. Diabetes Rev*., 2024. DOI: https://doi.org /10.2174/0115733 998296228240521 151 0 50.

[3] H. Alp, A. Sahin, P. Karabaglı, S. Karaburgu, B. Y. Sanal, and E. B. Yuksel, "Current perspective on diabetes mellitus in clinical sciences," *Nobel Tıp Bookstores*, 2023. DOI: https://doi.org/10.1677/JOE-09-0260.

[4] M. Sheikh et al., "Diabetes trends and their impact on public health," *Clin. Diabetes Endocrinol*, vol. 9, pp. 1–15, 2023. DOI: https://doi.org/10.1186/s40842-023-00163-5.

[5] N. S. Levitt, "Diabetes in Africa: Epidemiology, management, and healthcare challenges," *Heart*, vol. 94, no. 11, pp. 1376–1382, 2008. DOI: https://doi.org/10.1136/hrt.2008.147306.

[6] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat. Med*., vol. 28, no. 1, pp. 31–38, 2022. DOI: https://doi.org/10.1038/s41591-021-01614-0.

[7] A. S. Abdalkafor, N. M. Aiman, and O. N. Mustafa, "Predicting the success rates of schools using artificial neural networks," *J. Theor. Appl. Inf. Technol*., vol. 96, no. 19, pp. 6339–6348, 2018.

[8] R. S. Peres, X. Jia, J. Lee, K. Sun, A. W. Colombo, and J. Barata, "Industrial artificial intelligence in Industry 4.0: Systematic review, challenges, and outlook," *IEEE Access*, vol. 8, pp. 220121–220139, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.3042874.

[9] M. Wakchaure, B. K. Patle, and A. K. Mahindrakar, "Application of AI techniques and robotics in agriculture: A review," *Artif. Intell. Life Sci*., vol. 3, p. 100057, 2023. DOI: https://doi.org/10.1016/j.ailsci.2023.100057.

[10] L. Wang and M. Zhou, "Deep learning in agriculture: Challenges and opportunities," *Comput. Agric. Syst.*, vol. 12, no. 2, pp. 30–40, 2022. DOI: https://doi.org/10.1016/j.compag.2022.09.001.

[11] D. S. Joseph, P. M. Pawar, and K. Chakradeo, "Real-time plant disease dataset development and detection of plant disease using deep learning," *IEEE Access*, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3358333.

[12] M. Kong and S. J. Song, "Artificial intelligence applications in diabetic retinopathy: Current trends and future expectations," *Endocrinol. Metab. (Seoul, Korea)*, vol. 38, no. 3, pp. 195–204, 2023. DOI: https://doi.org/10.3803/EnM.2023.1913.

[13] Y. Li et al., "Multimodal information fusion for glaucoma and diabetic retinopathy classification," in *Proc. Int. Workshop Ophthalmic Med. Image Anal*., 2022, pp. 53–62.

[14] J. Chen and A. Gupta, "AI-assisted multimodal imaging in diabetic retinopathy detection," *Ophthalmic Sci. Rev*., vol. 14, no. 1, pp. 72–88, 2022. DOI: https://doi.org/10.1016/j.osr.2022.06.002.

[15] M. El Habib Daho et al., "Improved automatic diabetic retinopathy severity classification using deep multimodal fusion of UWF-CFP and OCTA images," in Proc. *Int. Workshop Ophthalmic Med. Image Anal*., 2023, pp. 11–20. DOI: https://doi.org/10.1007/978-3-031-44013-7_2.

[16] G. U. Nneji, J. Cai, J. Deng, H. N. Monday, M. A. Hossin, and S. Nahar, "Identification of diabetic retinopathy using weighted fusion deep learning based on dual-channel fundus scans," *Diagnostics*, vol. 12, no. 2, p. 540, 2022. DOI: https://doi.org/10.3390/diagnostics12020540.

[17] B. Fetaji, M. Fetaji, M. Ebibi, and M. Ali, "Predicting diabetes using machine learning algorithms: Comparative analysis," in Proc. Int. Conf. Emerg. Technol. *Comput.*, 2021, pp. 185–193. DOI: https://doi.org/10.1007/978-3-030-90016-8_13.

[18] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci*., vol. 132, pp. 1578–1585, 2018. DOI: https://doi.org/10.1016/j.procs.2018.05.122.

[19] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat*., vol. 2, no. 4, pp. 433–459, 2010. DOI: https://doi.org/10.1002/wics.101.

[20] M. J. Masnan et al., "Principal component analysis—A realization of classification success in multi-sensor data fusion," Princ. Compon. Anal. Appl., pp. 1–25, 2012. DOI: https://doi.org/10.13140/2.1.2807.3926.

[21] H. Alshaher, "Studying the effects of feature scaling in machine learning," *North Carolina Agric. Tech. State Univ.*, 2021. DOI: https://doi.org/10.5555/AAI28772109.

[22] P. J. M. Ali, R. H. Faraj, E. Koya, and P. J. M. Ali, "Data normalization and standardization: A technical report," *Mach. Learn. Tech. Rep*., vol. 1, no. 1, pp. 1–6, 2014. DOI: https://doi.org/10.13140/RG.2.2.28948.04489.

[23] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Compute. Electr. Eng*., vol. 40, no. 1, pp. 16–28, 2014. DOI: https://doi.org/10.1016/j.compeleceng.2013.11.024.

[24] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Syst. Appl*., vol. 41, no. 14, pp. 6371–6385, 2014. DOI: https://doi.org/10.1016/j.eswa.2014.04.019.

[25] M. Mareli and B. Twala, "An adaptive cuckoo search algorithm for optimization," *Appl. Compute. Informatics*, vol. 14, no. 2, pp. 107–115, 2018. DOI: https://doi.org/10.1016/j.aci.2017.09.001.

[26] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of the nearest neighbour algorithm for learning and classification," in *Proc. Int. Conf. Intell. Compute. Control Syst*. (ICCS), 2019, pp. 1255–1260. DOI: https://doi.org/10.1109/ICCS45141.2019.9065747.

[27] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowl.-Based Syst*., vol. 60, pp. 20–27, 2014. DOI: https://doi.org/10.1016/j.knosys.2013.12.023.

[28] S. A. Rafa et al., "Bird species detection utilizing an effective hybrid model," in Proc. *Int. Multi-Conf. Syst. Signals Devices (SSD)*, 2024, pp. 705–710. DOI: https://doi.org/10.1109 /SSD61 670. 2024. 105 49480.