Research Article

Penikalapati Pragathi and Agastyaraju Nagaraja Rao*

An effective integrated machine learning approach for detecting diabetic retinopathy

https://doi.org/10.1515/comp-2020-0222
received July 27, 2020; accepted January 11, 2021

Abstract: Millions of people across the world are suffering from diabetic retinopathy. This disease majorly affects the retina of the eye, and if not identified priorly causes permanent blindness. Hence, detecting diabetic retinopathy at an early stage is very important to safeguard people from blindness. Several machine learning (ML) algorithms are implemented on the dataset of diabetic retinopathy available in the UCI ML repository to detect the symptoms of diabetic retinopathy. But, most of those algorithms are implemented individually. Hence, this article proposes an effective integrated ML approach that uses the support vector machine (SVM), principal component analysis (PCA), and moth-flame optimization techniques. Initially, the ML algorithms decision tree (DT), SVM, random forest (RF), and Naïve Bayes (NB) are applied to the diabetic retinopathy dataset. Among these, the SVM algorithm is outperformed with an average of 76.96% performance. Later, all the aforementioned ML algorithms are implemented by integrating the PCA technique to reduce the dimensions of the dataset. After integrating PCA, it is noticed that the performance of the algorithms NB, RF, and SVM is reduced dramatically; on the contrary, the performance of DT is increased. To improve the performance of ML algorithms, the moth-flame optimization technique is integrated with SVM and PCA. This proposed approach is outperformed with an average of 85.61% performance among all the other considered ML algorithms, and the classification of class labels is achieved correctly.

Keywords: diabetic retinopathy, support vector machine, machine learning, moth-flame optimization, classification, measures, principal component analysis

1 Introduction

Several classification techniques of machine learning (ML) algorithms were discussed, and these techniques greatly helped the stakeholders of the medical field for predicting heart disease. A model of the artificial neural network (ANN) was proposed by Dangare and Apte was outperformed with 100% accuracy [1]. To detect anomalies in hyperglycemia classification, techniques such as feedforward ANN, deep belief network, genetic algorithm (GA), support vector machine (SVM), and Bayesian neural network were proposed and implemented [2]. Some of the ML techniques were rarely implemented or not implemented at all. Besides, the accuracy of some ML techniques was lower than the accuracy obtained by DL techniques. Hence, the combined models of ML and deep learning (DL) techniques were discussed to enhance accuracy for diabetes prediction [3]. The principal component analysis (PCA) was discussed to deal with large datasets. The dimensions of these large datasets could be reduced by using PCA to observe the correlation between the attributes and for better interpretability [4]. Diabetic retinopathy would affect eyes. The current trends of disease, mechanisms, and approaches to treat diabetic retinopathy were discussed [5]. A system with ML algorithms namely, k-nearest neighbor (KNN), variants of SVM, and NB was discussed to detect exudates in retinal images automatically. The proposed system detected the exudates with an accuracy of 98.58% that was greater than other existing techniques [6]. The classification techniques such as C4.5, Naïve Bayes (NB), and clustering technique k-means clustering were used to detect the risk factors of diabetes disease complications. The proposed system achieved an average accuracy of 68% [7]. Moth-flame optimization algorithm was discussed, which would improve the accuracy of classification [8]. To detect and classify characteristics such as micro-aneurysms (MA) and hemorrhages in retinal images, a model with convolutional neural networks (CNN) was proposed. The proposed model achieved 95% accuracy for the two-class classification of the dataset size 30,000 images and 85% for the

* **Corresponding author: Agastyaraju Nagaraja Rao,** School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India, e-mail: nagarajaraoa@vit.ac.in
Penikalapati Pragathi: School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India, e-mail: pragathi.2016@vit.ac.in

five-class classification of the dataset size 3,000 images [9]. Diseases related to heart, breast cancer, and diabetes were analyzed using ML techniques. This study revealed the significance of predicting the risk factors of diseases [10]. The aforementioned discussions exhibit the role of ML in predicting the symptoms and risk factors of different kinds of chronic diseases. At this moment, the statement "prevention is better than cure" is to be reminded. If the symptoms of the disease are identified before the occurrence, then it will help the people to take necessary precautions. Hence, the ML algorithms have great participation and impact on medical diagnosis.

1.1 State-of-the art literature review

The aforementioned works represent various ML algorithms and their corresponding accuracies in the medical diagnosis. In this section, in addition to mentioned earlier, some more state-of-the-art related works are presented as follows: PCA-based techniques were discussed in refs [11–14], where PCA and K-means techniques were integrated with logistic regression for predicting diabetes [11], PCA and linear discriminant analysis (LDA) were discussed for reducing dimensions of a large dataset cardiotocography [12], a deep neural network based on the PCA-firefly method was proposed to detect the signs of diabetic retinopathy at an early stage [13], and PCA-firefly-based classification model with the XGBoost classification method was discussed [14]. SVM-based techniques were discussed in refs [15,16], where SVM and simulated annealing (SA) were proposed for diagnosing the disease hepatitis [15], and SVM with a fruit fly optimization algorithm was proposed to classify medical data effectively [16]. Neural network-based approaches were discussed in refs [17,18], where a multilayer perceptron NN with backpropagation was selected to develop a system that predicts the risk factors of heart disease [17], and a model of deep CNN was proposed to notice and classify the diabetic retinopathy in retinal images [18]. ML algorithms were discussed in refs [19–21], where ANN, K-means clustering, and random forest (RF) algorithms were proposed and implemented for predicting diabetes early. Among these algorithms, the ANN outperformed with an accuracy of 75.7% [19], the techniques such as DT, SVM, LDA, and NB. were implemented. The LDA performed well with an accuracy of 79% including hypertension and prehypertension [20], and a classification model was proposed using the techniques such as SVM, NB, KNN, and DT for predicting diabetes [21]. DL techniques were discussed in refs [22,24], where a customized deep CNN was used in the proposed model to automate the fundus images' classification for detecting the diabetic retinopathy [22], and ensemble models of deep CNN such as Dense121, Resnet50, Dense169, Xception, and Inceptionv3 were implemented for detecting diabetic retinopathy [23], and a deep CNN model was proposed to classify fundus image and for the grading of Macular Edema [24]. GA-based approaches were discussed in refs [25–27], where an SVM classifier was used for dual classification, and later, these results are combined and fed into a GA to detect diabetic retinopathy [25], a GA- and SVM-based approaches were proposed to diagnose heart disease [26], and a hybrid GA and fuzzy logic classifier were proposed for diagnosing heart disease [27]. An ensemble-based approach were proposed for automated diagnosis and screening of diabetic retinopathy. The proposed approach provided higher accuracy [28]. A moth-flame optimization algorithm were proposed, and the performance was compared with other nature-inspired algorithms [29]. A hybrid firefly-bat optimized fuzzy ANN classifier was proposed for predicting diabetes, and it performed well than other convolutional methods [30]. A hybrid metaheuristic algorithm was proposed by techniques such as whale optimization algorithm and SA [31]. The combination of the elemental analysis of diabetic toenails and ML approaches was proposed to classify type-2 diabetes [32]. Cox proportional hazard, a regression-based method, was implemented to detect cardiovascular disease at an early stage [33]. A model was proposed for predicting the risk of gestational diabetes [34]. An RF classifier was used for predictions in the proposed algorithm DMP-MI to classify diabetes mellitus [35]. The aforementioned works show the wider implementation of various ML algorithms in medical diagnosis. It is observed that most of the previous works were carried out with the prime focus on a performance measure "accuracy" to evaluate the performance of a classifier. This article proposes an integration of SVM, PCA, and moth-flame optimization techniques for predicting the class labels of diabetic retinopathy. The proposed integrated approach evaluates the performance of a classifier using the measure "accuracy" as the same as in the previous works. Besides, measures such as sensitivity, recall, specificity, precision, and $F1$-score are also used. The proposed model contributes to a comprehensive analysis of ML algorithms' performance for the aforementioned measures and the classification of class labels.

2 Proposed methodology and implementation

To implement the proposed methodology, the dataset diabetic retinopathy is retrieved from the UCI ML repository. The proposed techniques such as normalization, PCA, and SVM are briefly described as follows. The dimensions of a dataset may consist of different levels of data. If the dataset is directly taken for the computation process, the higher-ordered dimensions may dominate the other lower-ordered dimensions. The obtained result is no way useful for decision-making. So, it is necessary to scale the data to make all the dimensions to be at the same level. The normalization technique scales the data and removes anomalies existing in the data. This is a preprocessing technique generally applied to the dataset before proceeding with the analysis process. The normalized data usually lies between 0 and 1. The main functionality of the PCA is to reduce the dimensions of a dataset. This is also referred to as dimensionality reduction. When a dataset contains more dimensions, several dimensions might be highly correlated. This problem is referred to as multicollinearity. The existence of multicollinearity affects the quality of data analysis. The PCA technique is best to deal with the multicollinearity problem. The main elements in PCA are principal components (PCs). The generation of the number of PCs is based on the number of dimensions given as input. The same number of PCs will be generated for the given number of dimensions, and they are ordered by their variance. PCs with high variance come first and then next higher level and so on. SVM is one of the most widely used supervised ML techniques. It is largely used for classification in high-dimensional data. It can be applied to both linearly and non-linearly separable problems. The key elements of the SVM are hyperplanes. In SVM, the classification of data will be done by identifying the hyperplanes. Support vectors are the vectors that describe the hyperplanes. In Section 2.1, the description of the dataset is given. In Section 2.2, the algorithm and objective function for the moth-flame optimization technique is described. The flow of activities in the proposed model is shown in Figure 1.

As an initial step, the collected diabetic retinopathy dataset is inputted to the proposed model. The details of the retinal images presented in this dataset are used to classify the images and to decide whether there are any symptoms of the diabetic retinopathy existing. Before proceeding to implement ML algorithms, it is important to normalize the data. This normalization can be done by using a standard scaler. After normalization, the ML algorithms such as DT, NB, RF, and SVM will be individually applied to the dataset. Next, the PCA technique is implemented for reducing the features in the dataset. This feature reduction will enable all the features to be at the same level. It means that the domination of higher-order features will be avoided. Now, the reduced features are inputted into aforesaid ML algorithms. The performance of ML algorithms before and after the implementation of PCA is observed. If high performance is observed, then we can continue with the recently applied technique. Otherwise, to improve the performance of ML algorithms, the resultants of previously applied techniques are feeded into the proposed integrated approach, i.e., SVM + PCA + moth-flame optimization technique. Then, the performance of ML algorithms was evaluated and compared concerning the performance measures discussed in Section 1.1. By looking into the comparative analysis, it can be understood that the ML approach that outperformed among all the others represents the correct classification of class labels.

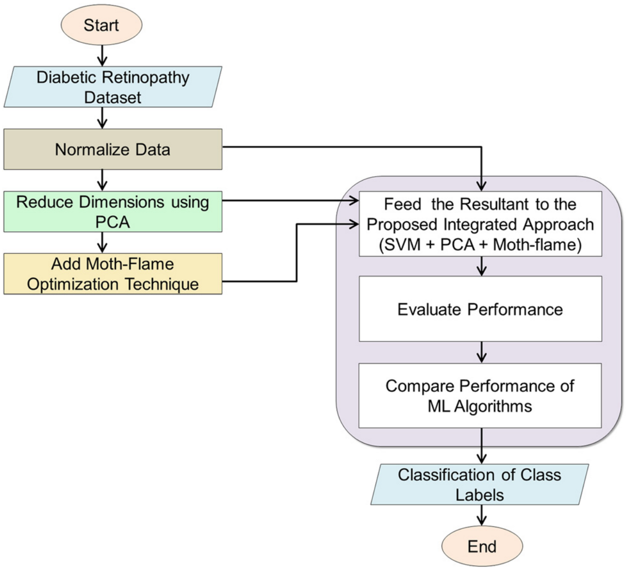
Start
Diabetic Retinopathy Dataset
Normalize Data
Reduce Dimensions using PCA
Add Moth-Flame Optimization Technique
Feed the Resultant to the Proposed Integrated Approach (SVM + PCA + Moth-flame)
Evaluate Performance
Compare Performance of ML Algorithms
Classification of Class Labels
End

Figure 1: Flowchart for the proposed integrated ML approach.

2.1 Dataset description

The diabetic retinopathy dataset comprises 20 features that represent the Messidor image set. Extracted features from an image will reveal the existence and non-existence of diabetic retinopathy. Features represented in the dataset will provide the information on any detected injury/lesion or description of the image. All 20 features are numbered from 0 through 19 as presented in Table 1.

Table 1: Description of dataset features

Feature number	Description of feature
0	Image quality is represented as *binary values* 1 and 0. 1 = good quality, and 0 = bad quality
1	Pre-screening information is represented as *binary values* 1 and 0. 1 = severe abnormality in the retina, and 0 = no abnormality
2–7	These features represent the number of MA values detected. MA causes retinal blood leakage. These features show the results at confidence intervals 0.5 through 1 respectively
8–15	Same as 2–7 for exudates. These are normalized to make all the features at the same level
16	This feature gives Euclidean distance information between the centers of the macula and optic disc
17	This feature contains information about optic disc diameter
18	The binary values of classification based on amplitude modulation (AM) and frequency modulation (FM)
19	Class labels are represented as *binary values* 1 and 0. 1 = symptoms of diabetic retinopathy, and 0 = no symptoms

In the dataset, feature 0 consists of the values related to the quality of the image. If lesions are identified effectively in the captured image, it is said to have good quality otherwise bad quality. This quality is represented with binary values 1 and 0. The existence of value 1 means, the image contains good quality, and the value 0 means, the image contains bad quality. Feature 1 consists of the details of pre-screening. This feature is also described with binary values 1 and 0. The existence of value 1 represents that there is a severe abnormality and 0 represents no abnormality in the retina. Features from 2 to 7 consist of the number of values detected that is related to microaneurysms (MA). The microaneurysms cause blood leakage to the tissues of the retina. These MA values are detected with confidence intervals from 0.5 through 1. Features from 8 to 15 consist of the normalized values related to exudates. These are represented as same as the features from 2 to 7. The normalization makes all the features to be at the same level. Feature 16 gives the Euclidean distance information between the centers of the macula and optic disc. Feature 17 consists of the details of the optic disc diameter. The starting point of retinal blood vessels is an optic disc. Feature 18 consists of the binary values related to the classification based on the modulations AM and FM. Finally, feature 19 consists of the binary values related to class labels. The value 1 represents the existence of symptoms of diabetic retinopathy, and 0 represents no symptoms existing.

2.2 Proposed algorithm

The general phenomenon of the moth-flame optimization technique is described as follows. The moth-flame optimization technique also referred to as a population-based technique. In this technique, both moths and flames are said to be solutions. The moths are said to be agents of search space, and flames are said to be the best positions. The difference between these two depends on the update of each iteration. By updating the position at each iteration, the moth never misses the best position. The steps of the proposed technique are given in Algorithm 1.

Algorithm 1 Algorithm for moth-flame optimization technique

```
1: Initiate the parameters.
2: Initiate the generation of moths randomly.
3: Identify the fitness functions and mark the best
   positions of flames.
4: Update flame numbers.
5: Calculate distance related to moth.
6: Update the positions related to moth.
7: Repeat the steps 2–6 until the expected criteria
   achieved.
8: If criteria are achieved report the best positions of
   moths.
```

2.3 Objective function

The objective function of moth-flame is given in equations (1)–(11) as follows: initialization of moths is represented in a matrix as shown in (1):

$$M = \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & \dots & b_{1,q} \\ b_{2,1} & b_{2,2} & \dots & \dots & b_{2,q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{p,1} & b_{p,2} & \dots & \dots & b_{p,q} \end{bmatrix}, \tag{1}$$

where p = total number of moths and q = total number of variables. The fitness function for moths is given in equation (2):

$$\mathrm{FM} = \begin{bmatrix} \mathrm{FM}_1 \\ \mathrm{FM}_2 \\ \cdots \\ \mathrm{FM}_q \end{bmatrix}. \quad (2)$$

Initialization of flames is represented in a matrix ss shown in equation (3):

$$N = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & \cdots & c_{1,q} \\ c_{2,1} & c_{2,2} & \cdots & \cdots & c_{2,q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{p,1} & c_{p,2} & \cdots & \cdots & c_{p,q} \end{bmatrix}, \quad (3)$$

where p = total number of flames and q = total number of variables. The fitness function for flames is given in equation (4):

$$\mathrm{FN} = \begin{bmatrix} \mathrm{FN}_1 \\ \mathrm{FN}_2 \\ \cdots \\ \mathrm{FN}_q \end{bmatrix}. \quad (4)$$

The mathematical model of moth-flame optimization technique represented as a three-tuple is given in equation (5):

$$\mathrm{MNF} = (G, H, T). \quad (5)$$

G represents the random population of moths and fitness values is given in equation (6):

$$G = \phi \rightarrow \{M, \mathrm{FM}\}. \quad (6)$$

The function H decides the moth movement for finding the best position of flame and updates every time, which is expressed in equation (7):

$$H = M \rightarrow M. \quad (7)$$

The function F determines whether it is true or false:

$$F = M \rightarrow \{\text{true, false}\}. \quad (8)$$

The equation for updating the position is given in equation (9):

$$M_k = S(M_j, N_k), \quad (9)$$

where S = spiral function, M_j = jth moth, and N_k = kth flame. The spiral path of the moth flow logarithmically is given in (10):

$$S(M_k, N_l) = N_k \cdot \mathrm{e}^{wc} \cdot \cos(2\pi c) + N_l, \quad (10)$$

where w = constant and c values lie between −1 and 1. Calculation of distance between the kth moth and lth flame is given in equation (11):

$$\begin{cases} D = |N_l - M_j| \\ c = z - 1 * (\text{rand} + 1), \end{cases} \quad (11)$$

where z value varies between −1 and −2, and when z value is less, it represents that the moth is closer to the flame.

3 Results and discussion

The implementation of ML algorithms was performed on a diabetic retinopathy dataset retrieved from the UCI ML repository. The performance of the proposed integrated ML model is evaluated using measures such as precision, $F1$-score, specificity, accuracy, recall, and sensitivity (12)–(16). Precision describes the correctness and is given in equation (12). Recall/sensitivity represents the wholeness and is expressed in equation (13). $F1$-score is described in equation (14). Accuracy characterizes the rightness and is given in equation (15). Specificity is described in equation (16).

$$\text{Precision} = \frac{\mathrm{PT}}{\mathrm{PF} + \mathrm{PT}}, \quad (12)$$

$$\text{Recall/sensitivity} = \frac{\mathrm{PT}}{\mathrm{NF} + \mathrm{PT}}, \quad (13)$$

$$F1\text{-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (14)$$

$$\text{Accuracy} = \frac{\mathrm{PT} + \mathrm{NT}}{\mathrm{PF} + \mathrm{PT} + \mathrm{NF} + \mathrm{NT}}, \quad (15)$$

$$\text{Specificity} = \frac{\mathrm{NT}}{\mathrm{PF} + \mathrm{NT}}, \quad (16)$$

where PT = positive (true value), PF = positive (false value), NT = negative (true value), and NF = negative (false value). The simulation results of ML algorithms and corresponding performance measures are described as follows: First, the dataset has experimented with four popular ML algorithms, namely, DT, NB, RF, and SVM. Figure 2 depicts the performance of these algorithms.

From Figure 2, it can be observed that the DT classifier has achieved 57% of precision, recall, and $F1$-score, and 57.1, 54.5, and 59.2% of accuracy, sensitivity, and specificity, respectively. The NB classifier has achieved 64% of precision, 63% of recall and $F1$-score, and 63.2, 64.2, and 62.4% of accuracy, sensitivity, and specificity, respectively. The RF classifier has achieved 70% of precision, recall, and 69% of $F1$-score and 68.8, 76.2, and 63% of accuracy, sensitivity, and specificity, respectively. The algorithm SVM has achieved 79% of precision, 76%

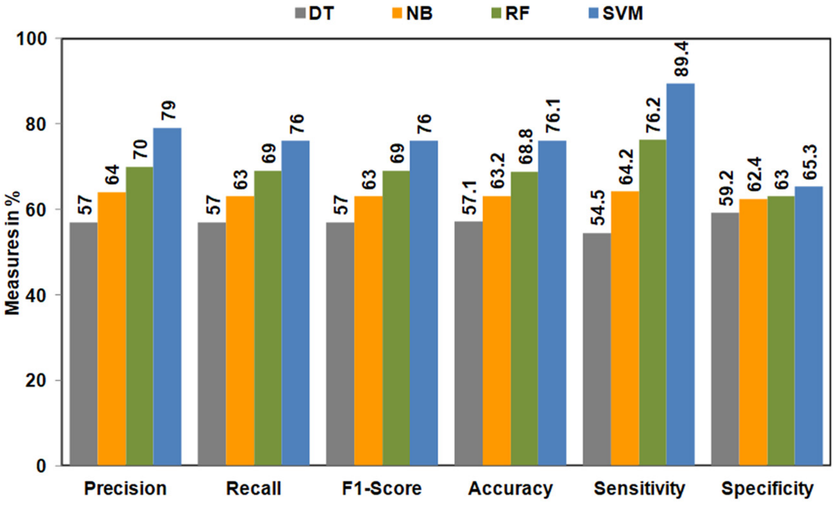

Figure 2: Performance of ML algorithms on the original dataset.

of recall and *F*1-score, and 76.1, 89.4, and 65.3% of accuracy, sensitivity, and specificity, respectively. When these performances are compared with each other, it is observed that the SVM is outperformed.

The dataset is then fed in to the PCA for dimensionality reduction. The PCA has reduced the dataset from 20 dimensions to 12 dimensions. These reduced features of the dataset are given as input to the aforementioned classifiers. Figure 3 depicts the results obtained after applying ML algorithms on reduced dimensions.

From Figure 3, it can be observed that the DT classifier has achieved 67% of precision, recall, and *F*1-score, and 67.09, 66.3, and 67.6% of accuracy, sensitivity, and specificity, respectively. The NB classifier has achieved 61% of precision, 60% of recall and *F*1-score, and 59.7% of accuracy, sensitivity, and specificity. The RF classifier has achieved 69% of precision, recall, and *F*1-score, and 69.2, 67.8, and 70.4% of accuracy, sensitivity, and specificity, respectively. The SVM has achieved 71% of precision, 66% of recall and *F*1-score, and 65.8, 81.3, and 55.7% of accuracy, sensitivity, and specificity, respectively. Figure 3 shows that the performance of the classifiers NB, RF, and SVM has been reduced when PCA is applied. But, the performance of DT has been enhanced with PCA. The main reason for the degradation of performance is dimensionality reduction. It is understood

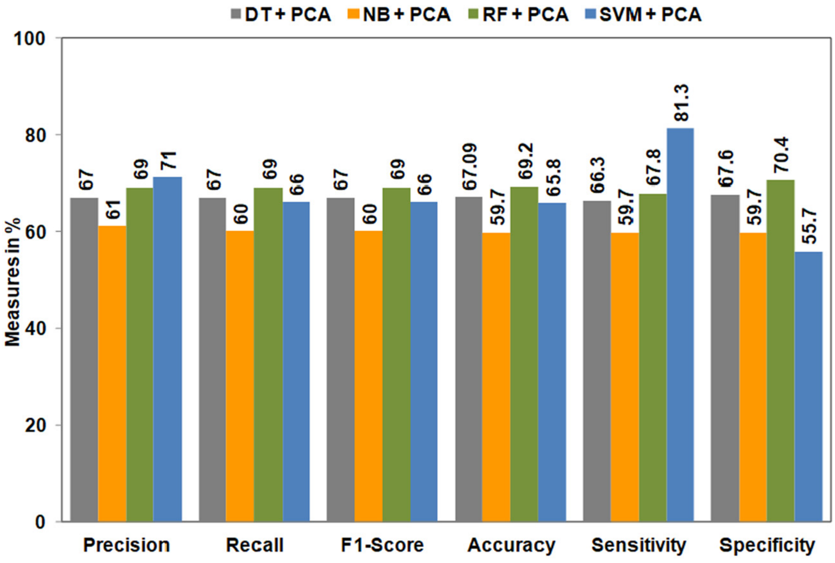

Figure 3: Performance of ML algorithms after reducing dimensions using PCA.

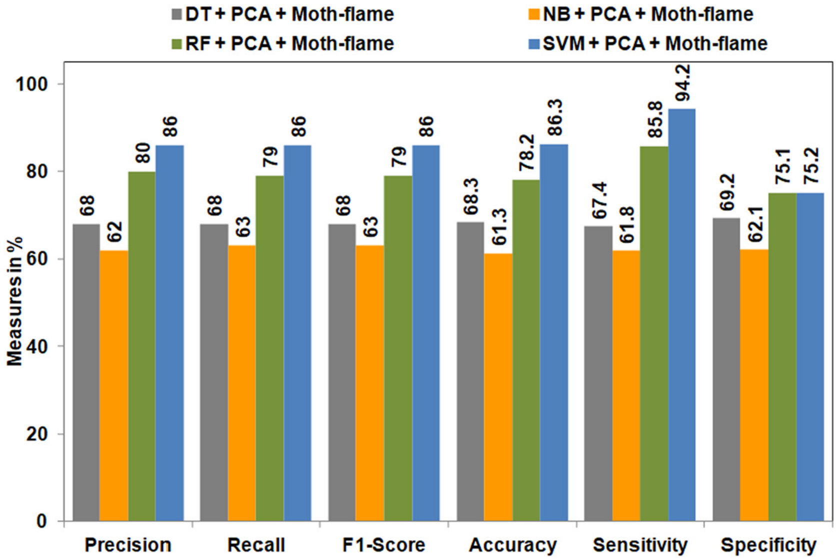

Figure 4: Performance of ML algorithms after implementation of moth-flame optimization technique.

that these classifiers work better with more samples and dimensions.

To achieve better performance of PCA-based classifiers, the moth-flame optimization technique is applied. This technique eliminated the attributes that are affecting the performance of the classification model negatively. It chooses the optimal features/attributes that have a positive impact on performance. By implementing the moth-flame technique, the 12 dimensions were further reduced to 9. Based on this reduction, it is understood three dimensions are affecting the performance of the model. These three dimensions are eliminated by the moth-flame algorithm. Figure 4 depicts the result of the model with PCA and moth-flame optimization techniques.

Figure 4 shows that the DT classifier has achieved 68% of precision, recall, and $F1$-score, and 68.3, 67.4, and 69.2% of accuracy, sensitivity, and specificity, respectively. The NB classifier has achieved 62% of precision, 63% of recall and $F1$-score, and 61.8% of accuracy, sensitivity, and specificity is 62.1%. The RF classifier has achieved 80% of precision, 79% of recall, and $F1$-score, and 78.2, 85.8, 75.1% of accuracy, sensitivity, specificity, respectively. The SVM has achieved 86% of precision, recall, and $F1$-score, and 86.3, 94.2, and 75.2% of accuracy, sensitivity, and specificity, respectively. From Figure 4, it is evident that the performance of all the classifiers except NB has improved dramatically with the addition of the moth-flame optimization technique. This technique eliminated the dimensions that impact the performance of the classifiers negatively. It is observed from the proposed model that the integration of SVM, PCA, moth-flame outperformed than all other ML algorithms. The performance of all ML algorithms is summarized in Table 2. In the table, PC, MF, Prscn, Rcl, $F1$, Acc, Sens, Spec refer to priniciple component analysis, moth-flame, precision, recall, $F1$-score, accuracy, sensitivity and specificity respectively. It can be noticed that the proposed integrated model of SVM, PCA, moth-flame has achieved high performance than the other ML algorithms. The high performance of this model represents that the classification of class labels has been achieved correctly.

Table 2: Performance summary of ML algorithms

ML algorithms	Prscn	Rcl	$F1$	Acc	Sens	Spec
DT	57	57	57	57.1	54.5	59.2
NB	64	63	63	63.2	64.2	62.4
RF	70	69	69	68.8	76.2	63
SVM	79	76	76	76.1	89.4	65.3
DT + PC	67	67	67	67.09	66.3	67.6
NB + PC	61	60	60	59.7	59.7	59.7
RF + PC	69	69	69	69.2	67.8	70.4
SVM + PC	71	66	66	65.8	81.3	55.7
DT + PC + MF	68	68	68	68.3	67.4	69.2
NB + PC + MF	62	63	63	61.3	61.8	62.1
RF + PC + MF	80	79	79	78.2	85.8	75.1
SVM + PC + MF	86	86	86	86.3	94.2	75.2

4 Conclusion

To identify the diabetic retinopathy, this article proposed an integrated approach of ML algorithms and achieved high performance. The dataset that is retrieved from the UCI ML repository is used for the proposed approach. The key observations of the proposed integrated approach are as follows:

- From Figure 2, it is observed that when the ML algorithms are implemented individually, SVM is outperformed than other ML algorithms.
- From Figure 3, it is evident that the reduction of the dimensions using the PCA technique has negatively influenced the performance of majority ML algorithms.
- From Figure 4, it is understood that the integration of SVM, PCA, and moth-flame optimization techniques improved the performance of classification and identified the class labels correctly.
- From Table 2, it is easy to interpret and compare the performances achieved by the implemented ML algorithms.

Hence, the proposed integrated approach of ML algorithms is very useful for detecting diabetic retinopathy to prevent blindness.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: The datasets generated during and/or analyzed during the current study are available in the Diabetic Retinopathy Debrecen Data Set, https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set.

References

[1] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *Int J Comput Appl.*, vol. 47, no. 10, pp. 44–48, 2012.

[2] J. D. Elia, J. K. Sun, and W. S. Alan, "Diabetic retinopathy: current understanding mechanisms, and treatment strategies," *JCI Insight*, vol. 2, pp. 1–13, 2017.

[3] M. I. Al-janabi, M. H. Qutqut, and M. Hijjawi, "Machine learning classification techniques for heart disease prediction: a review," *Int J Eng Technol.*, vol. 7, pp. 5373–5379, 2018.

[4] A. W. Zebene, A. Eirik, T. Botsis, D. Albers, M. Lena, and H. Gunnar, "Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes," *J Med Internet Res.*, vol. 21, pp. 1–18, 2019.

[5] S. L. M. Sainte, A. Linah, A. Rana, and T. Saba, "Current techniques of diabetes prediction: review and case study," *Appl. Sci.*, vol. 9, pp. 1–19, 2019.

[6] A. Javeria, M. Sharif, M. Yasmin, H. Ali, and S. F. Lawrence, "A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions," *J. Comput. Sci.*, vol. 19, pp. 153–164, 2017.

[7] F. Cut, E. M. Sipayung, and M. Siti, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Proc. Comput. Sci.*, vol. 161, pp. 449–457, 2019.

[8] G. Sumalatha and N. J. R. Muniraj, "Survey on medical diagnosis using data mining techniques," *International Conference on Optical Imaging Sensor and Security*, Coimbatore, India, 2013.

[9] R. Ghosh, G. Kuntal, and S. Maitra, "Automatic detection and classification of diabetic retinopathy stages using CNN," *International Conference on Signal Processing and Integrated Networks*, Noida, India, 2017.

[10] A. E. Ahmed, A. T. Sahlol, and A. A. Mohamed, "A bio-inspired Moth-flame optimization algorithm for Arabic handwritten letter recognition," *International Conference on Control Artificial Intelligence, Robotics & Optimization*, Prague, Czech Republic, 2017.

[11] C. Zhu, C. I. Uwa, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatic Med Unlocked*, vol. 17, pp. 1–7, 2019.

[12] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[13] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I. H. Ra, et al., "Early detection of diabetic retinopathy using PCA-firefly-based deep learning model," *Electronics*, vol. 9, pp. 1–16, 2020.

[14] S. Bhattacharya, S. R. K. S, P. K. R. Maddikunta, R. Kaluri, S. Singh, T. R. Gadekallu, et al., "A novel PCA-firefly-based XGBoost classification model for intrusion detection in networks using GPU," *Electronics*, vol. 9, pp. 1–16, 2020.

[15] J. S. Salimi, M. Z. Hossein, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Comput. Meth. Prog. Bio.*, vol. 108, pp. 570–579, 2012.

[16] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, et al., "Evolving support vector machines using fruit fly optimization for medical data classification," *Knowledge-Based Syst.*, vol. 96, pp. 61–75, 2016.

[17] S. Poornima, S. Singh, and G. S. J. Pandi, "Effective heart disease prediction system using data mining techniques," *Int. J. Nanomed.*, vol. 13, pp. 121–124, 2018.

[18] H. D. Jude, D. Omer, and U. Kose, "An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network," *Intell. Biomed. Data Anal. Process.*, vol. 32, pp. 707–721, 2019.

[19] T. Mahboob Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. Imtiaz Baig, et al., "A model for early prediction of diabetes," *Informatics Med. Unlocked*, vol. 16, pp. 1–6, 2019.

[20] H. Chirath and C. Charith, "A Machine learning approach to predict diabetes using short recorded photoplethysmography and physiological characteristics," *Artif. Intell. Med.*, vol. 11526, pp. 322–327, 2019.

[21] W. Mitesh, V. Kumar, S. Tarale, G. Payal, and D. J. Chaudhari, "Diabetes diagnosis using machine learning

algorithms," *Int. Res. J. Eng. Technol.*, vol. 6, pp. 1470–1476, 2019.

[22] G. Rishab and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, pp. 962–969, 2017.

[23] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, et al., "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530–150539, 2019.

[24] J. Sahlsten, J. Jaskari, J. Kivinen, L. Turunen, E. Jaanio, K. Hietala, et al., "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Sci. Rep.*, vol. 9, pp. 1–11, 2019.

[25] R. A. Welikala, M. M. Fraz, J. Dehmeshki, A. Hoppe, V. Tah, S. Mann, et al., "Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy," *Computerize. Med. Imag. Graphic.*, vol. 43, pp. 64–77, 2015.

[26] C. G. Babu and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Comput.*, vol. 22, pp. 14777–14787, 2019.

[27] T. G. Reddy, M. K. R. Praveen, L. Kuruva, R. D. Singh, K. Rajesh, and S. Gautam, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evolut. Intell.*, vol. 13, pp. 185–196, 2019.

[28] B. Antal and H. Andras, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-Based Syst.*, vol. 60, pp. 20–27, 2014.

[29] M. Seyedali, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," *Knowledge-Based Syst.*, vol. 89, pp. 228–249, 2015.

[30] R. G. Thippa and N. Khare, "Hybrid firefly-bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis," *Int. J. Intell. Eng. Syst.*, vol. 10, pp. 18–27, 2017.

[31] C. Iwendi, P. K. R. Maddikunta, G. T. Reddy, L. Kuruva, B. K. Ali, and M. P. Jalil, "A metaheuristic optimization approach for energy efficiency in the IoT networks," *Softw: Pract Exper*, vol. 51, pp. 1–14, 2020.

[32] C. A. Jake, C. S. Long, P. S. Beth, T. L. Smith, and L. D. George, "Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes," *Expert Syst. Appl.*, vol. 115, pp. 245–255, 2019.

[33] X. Jia, B. M. Mirja, M. Farhaan, and G. H. Hamid, "A cox-based risk prediction model for early detection of cardiovascular disease: Identification of key risk factors for the development of a 10-year CVD risk prediction," *Adv. Preventive Med.*, vol. 2019, pp. 1–11, 2019.

[34] B. M. Donovan, P. J. Breheny, J. G. Robinson, R. J. Baer, A. F. Saftlas, W. Bao, et al., "Development and validation of a clinical model for preconception and early pregnancy risk prediction of gestational diabetes mellitus in nulliparous women," *PLoS ONE*, vol. 14, pp. 1–21, 2019.

[35] Q. Wang, C. Weijia, J. Guo, J. Ren, C. Yongqiang, and D. N. Davis, "DMP-MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," *IEEE Access*, vol. 7, pp. 102232–102238, 2019.