International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

An Intelligent Harris Hawks Optimization (IHHO) based Pivotal Decision Tree (PDT) Machine Learning Model for Diabetes Prediction

[1]Roma Fayaz,[2]G.Vinoda Reddy, [3]M.Sujaritha, [4]N.Soundiraraj, [5]W.Gracy Theresa , [6]Dharmendra Kumar Roy, [7]J.Jeffin Gracewell (Corresponding Author),[8]S.Gopalakrishnan

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: In ancient times, an accurate diabetes prediction and type of classification are the most important and demanding tasks in the medical field for providing proper diagnosis to the patients. For this purpose, various machine learning based detection systems are developed in the conventional works to predict the diabetes from the given dataset. Still, it has some limitations with the factors of difficult to understand, high time requirement for training and testing, over fitting, and error outputs. Therefore, the proposed research work objects to implement a group of data mining techniques for developing an automated and efficient diabetes detection system. In this framework, an Inherent Coefficient Normalization (ICN) technique is implemented at first for preprocessing the PIMA Indian dataset obtained from the repository, which highly improves the quality of data for processing. Then, an Intelligent Harris Hawks Optimization (IHHO) technique is utilized to optimally select the features for training the classifier. Finally, the Pivotal Decision Tree (PDT) based classification technique is deployed to predict the data as whether diabetes or non-diabetes with reduced computational complexity and time consumption. During analysis, the performance and results of the proposed IHHO-PDT technique is validated and compared using various measures.

Keywords: *Diabetes Prediction, Machine Learning, Inherent Coefficient Normalization (ICN), Intelligent Harris Hawks Optimization (IHHO), Pivotal Decision Tree (PDT), and PIMA Indian Dataset.*

1. Introduction

Diabetes Mellitus (DM) [1, 2] is one of the most common and severe disease across many people over the countries. According to the world health report, it is analyzed that nearly 629 million of people can affect by this dangerous disease at 2045. Among other countries, it is rapidly increased in India and Sharan countries, and is very complex to predict that how much the disease is chronic and serious [3, 4]. Moreover, many of the medical organizations and healthcare industries object to control this disease for saving the people lives.

Specifically, it can affect the other body organs, so it is highly important to detect the disease at earlier stages for proper diagnosis and treatment.

Hence, the different types of prediction methodologies [5, 6] are developed in the conventional works for identifying diabetes based on the patients' information such as age, glucose level, blood sugar, pressure, BMI, and etc. The machine learning techniques are the most suitable options for developing an automated detection systems. Also, the classification techniques [7-10] are increasingly used in all kinds of applications for solving the complex problems. Then, it provides the predicted label for identifying that whether patient is affected by the disease or not.

The Support Vector Machine (SVM), Decision Tree (DT), Linear Regression (LR), Random Forest (RF), Neural Network (NN), Naïve Bayes (NB) and ensemble models [11-13] are the most commonly used machine learning approaches in the detection applications. Based on the existing review [14-16], it is analyzed that an optimization based classification techniques could provide an efficient performance in prediction and classification operations [17, 18]. The nature inspired and bio-inspired are the common optimization techniques used for providing solutions to solve the complex optimization problems. However, the existing

[1]Lecturer, Department of Computer Science, College of Computer Science and Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia, ***Email: rfayaz@jazanu.edu.sa***

[2]Professor, Department of Computer Science and Engineering (AI&ML), CMR Technical Campus, Kandlakoya, Medchal (M), Hyderabad, Telangana-501401, [3]Email:vinodareddy.cse@cmrtc.ac.in

*[3]Professor,Sri Krishna College of Engineering and Technology, Kuniyamuthur, Tamil Nadu 641008,India,****Email:sujaritham@skcet.ac.in***

[4]Assistant Professor Department of Electronics and Communication Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, 624622, India e-mail: soundar@psnacet.edu.in

[5]Associate Professor, Department of Computer and Science Engineering, Panimalar Engineering College, Chennai, Tamil Nadu 600123, India. ***Email: sunphin14@gmail.com***

[6]Assistant Professor, Department of Computer Science and Engineering, Hyderabad Institute of Technology and Management (HITAM), Hyderabad, Telangana, 501401,India ***Email:roy.dharmendra@gmail.com***

*[7]Assistant Professor, Department of Electronics and Communication, Saveetha Engineering College, Chennai, Tamil Nadu 602105,India,****Email:jgracewell02@gmail.com***

[8]Associate Professor, Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology (Deemed To be University),Chennai-600062, Tamil Nadu, India. Email: ***drsgk85@gmail.com***

works [19-22] faced many complications associated to the following factors: complex computational operations, difficult to understand, high time consumption, overfitting and error outputs. Therefore, the proposed work intends to implement a novel and intelligent optimization based classification model for developing an efficient diabetes prediction system. In this proposed framework, a combination of data mining techniques such as preprocessing, feature selection, and classification are deployed for diabetes prediction and classification. The major research objectives of this paper are as follows:

- To preprocess the given dataset by normalizing its attributes, an Inherent Coefficient Normalization (ICN) technique is implemented.
- To select the optimal features from the optimized feature set, an advanced Harris Hawks Optimization (HHO) technique is utilized, which supports to obtain an increased classification accuracy.
- To predict whether the given data is diabetes or not, an enhanced Pivotal Decision Tree (PDT) based classification methodology is deployed.
- To validate the performance and efficiency of proposed optimization based classification methodology, an extensive simulation has been carried out.

The other portions of this paper are structuralized into the following portions: Section II reviews the conventional optimization and classification methodologies related to diabetes prediction and detection. It also examines the benefits and demerits of each technique according to its features and working modules. Section III provides the explanation about the proposed IHHO based PDT classification technique used for developing an efficient diabetes detection system. Section IV validates the results of conventional and proposed techniques with respect to different evaluation parameters. Finally, the overall paper is concluded with its future work in Section V.

2. Related Works

This section reviews various machine learning based classification techniques used for predicting diabetes from the given dataset. It also discusses about the pros and cons of each machine learning methodology according to its key characteristics, and features.

Mujumdar, et al [23] intended to predict the Diabetes Mellitus (DM) disease by using an advanced machine learning approach. The key factor of this paper was to develop a new diabetes prediction model based on the factors of glucose, BMI, insulin and age. They examined about the performance of various machine learning techniques for predicting the diabetes from the given features. Here, the supervised, un-supervised, and semi-supervised based machine learning methodologies have been analyzed for developing an efficient diabetes prediction system. Also, it stated that the performance and efficacy of overall detection system was entirely depends on the quality of dataset. Hence, it should be enhanced with the use of machine learning based classification methodologies. According to this study, it was analyzed that an Ada Boost classifier outperforms the other techniques with performance results. Joshi, et al [24] objects to develop an efficient prediction model for earlier diagnosis of diabetes. In this work, the different machine learning classifiers such as Logistic Regression (LR), Artificial Neural Network (ANN), and Support Vector Machine (SVM) were validated and compared for identifying the most suitable technique for diabetes prediction. Typically, the classification was one of the most essential process in the detection systems, and it predicts the data as whether diabetic or non-diabetic. The key limitation of this work was, it failed to compute the evaluation measures to prove the improved results of classifiers.

Alehegn, et al [25] deployed an ensemble based classification methodology for predicting the diabetes from the PIMA dataset. Also, it targets to validate the performance of three different classification methodologies such as SVM, NB and Decision Stump (DS). This framework includes the modules of preprocessing, dataset training, prediction, and comparison. However, it limits with the problems of inefficient prediction, error outputs, and more time consumption. Abdulqadir, et al [26] utilized a recursive RF classification model for accurately detecting the diabetes in the healthcare systems. Here, the different types of data mining techniques have been used for improving the efficacy of overall prediction system. In this framework, the weighting parameter was estimated during feature selection, which helps to increase the accuracy of diabetes prediction. Yet, it has the problems of overfitting, not suitable for the multi-valued attributes, and inefficient detection rate. Battineni, et al [27] presented a comprehensive analysis for validating the performance of various machine learning techniques. The purpose of this work was to develop an efficient type-2 diabetes prediction system by using the data mining model. In addition to that, the k-cross fold validation method was also used in this work, which improvise the process of prediction. Arumugam, et al [28] deployed a multi-purpose machine learning methodology for predicting the multiple diseases associated to the DM. In this work, three different classifiers such as NB, DT, and SVM have been validated and compared in terms of error rate and accuracy. Based on this work, it was analyzed that the DT outperforms the other machine learning models with better detection results. Still, it faces the key problems of complex mathematical operations, lack of scalability, and reliability. The different types of data mining techniques used for developing the diabetes prediction system is shown in Fig 1.

Zou, et al [29] deployed a Principal Component Analysis (PCA) based machine learning classification methodology for predicting the DM. The key contribution of this paper was to minimize the number of features required for classification by using the combination of PCA and minimum Redundancy Maximum Relevance (mRMR) methodologies. Moreover, the RF based machine learning classifier was used to predict the DM from the given datasets based on the optimized features. The benefits of this work were increased accuracy, optimized performance and minimal time consumption. Garcia, et al [30] implemented a Sparse Auto Encoder (SAE) incorporated with the Convolutional Neural Network (CNN) for detecting diabetes from the given dataset. In this framework, the min-max normalization method was used to normalize the dataset for increasing the quality of data. Then, the data augmentation has been performed by using the Variational Auto Encoder (VAE), which helps to improve the detection performance. Finally, the CNN technique was used to predict the classified label as whether diabetes or non-diabetes. The primary advantages of this work were reduced error rate, minimal time requirement, and accurate results. Yet, it follows some mathematical operations for detection, which degrades the performance of entire system.

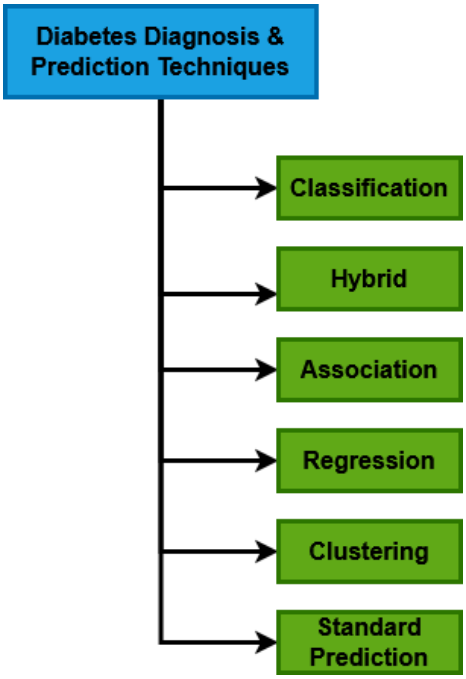

Fig 1. Types of data mining techniques used for diabetes prediction

Ramesh, et al [31] constructed a new remote healthcare monitoring framework for predicting the diabetes with the help of machine learning methodology. Here, the feature scaling and selection operations have been performed to increase the accuracy of classification. In addition to that, the data augmentation was applied before classification, which supports to reduce the error rate. Moreover, four different machine learning techniques such as KNN, LR, GNB, and SVM were utilized to validate the efficacy and performance of the classifier. Boutilier, et al [32] employed a machine learning based disease prediction model for identifying the diabetes from the given dataset. Here, the machine learning classification approach was mainly used to categorize both diabetes and hypertension based on various features like age, blood sugar, blood pressure, BMI, and etc. Abaker, et al [33] presented a comprehensive study for validating the performance of various machine learning approaches to select the suitable one for diabetes prediction and classification. In this work, the different types of feature selection approaches were also investigated for minimizing the dimensionality of features, which holds filter models, wrapper models, embedded models, ensemble models, and hybrid techniques. Mahmoud, et al [34] deployed a Hybrid Inductive Machine Learning Algorithm (HIMLA) for predicting diabetic retinopathy with reduced risk factors. Also, it used an optimization based classification methodology for ensuring the better system performance. However, it has the key problems of reduced accuracy, inefficient prediction, and lack of scalability.

From the literature, it is analyzed that the conventional works are mainly focused on developing an automated machine learning based prediction system for detecting diabetes. Still, it faced some complications associated to the factors of overfitting, high error rate, difficult to understand, complex mathematical operation, increased training and testing time. Therefore, the proposed work objects to implement a novel meta-heuristic optimization based classification mechanism for developing an efficient diabetes prediction system.

2. Proposed Methodology

This section presents the clear description about the proposed machine learning based classification methodology used for predicting the diabetes disease. The original contribution of this work is to implement a novel optimization integrated classification methodology to develop an efficient diabetes detection system. For this purpose, an intelligent Harris Hawks Optimization (HHO) based Pivotal Decision Tree (PDT) classification methodology is developed. Initially, the input dataset is obtained for processing, which is preprocessed by using an Inherent Coefficient Normalization (ICN) model for enhancing the quality of data. Typically, the overall prediction performance and efficiency of classifier are highly depends on the quality of input data. Hence, it should be highly enhanced before feature extraction and classification operations. Then, a HHO technique is deployed for optimally selecting the features from the preprocessed dataset, which supports to increase the detection accuracy and efficiency of classifier. Finally, the PDT technique is deployed to accurately predict the classified label based on the optimized set of features. The primary advantages of this work are as follows: reduced overfitting, minimal training & testing time, optimized performance, reduced complexity and high accuracy. The overall working flow of the proposed system is shown in Fig 2, which holds the following operations:

- Preprocessing based on Inherent Coefficient Normalization (ICN) Method
- Feature Selection using Harris Hawks Optimization (HHO)
- Disease Prediction using Pivotal Decision Tree (PDT) Classifier

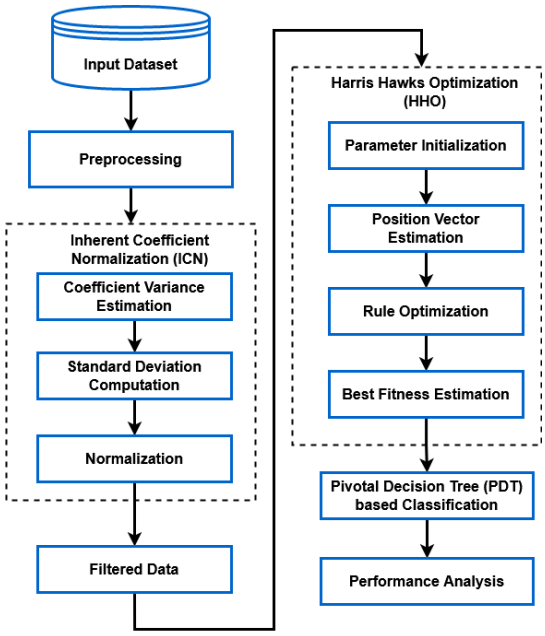

Fig 2. Working flow of the proposed system

3.1 Inherent Coefficient Normalization

At first, the input diabetes dataset is preprocessed for eliminating the noise and enhancing the quality, which is done by using an ICN technique. It is a kind of preprocessing approach mainly developed for normalizing the attributes of dataset. In this work, the PIMA Indian Diabetic dataset is used for analysis, which comprises the patient information related to the features of glucose level, BP, BMI, pedigree function, and etc. Generally, the

original dataset may have some irrelevant and incoherent information, which degrades the performance of recognition/prediction. Hence, the attributes are properly filtered and normalized before implementing further operations. In this preprocessing model, the coefficient estimation, standard deviation estimation, and normalization operations have been performed to generate the filtered data. The coefficient variance is computed by using the following model:

$$\mathrm{Co_V} = \frac{\mathrm{SD}}{\mathrm{Mean}} \tag{1}$$

Where, SD indicates the standard deviation, and $\mathrm{Co_V}$ is the coefficient variance. Then, the mean and SD values are computed as shown in below:

$$\mathrm{SD} = \sqrt{\frac{1}{\mathrm{x}-1}\sum_{\mathrm{d}=1}^{\mathrm{x}}(\mathrm{P_d} = \mathrm{Avg_V})^2} \tag{2}$$

$$\mathrm{Mean} = \frac{1}{\mathrm{X}}\sum_{\mathrm{d}=1}^{\mathrm{X}} \mathrm{P_d} \tag{3}$$

$$\boldsymbol{Nor_D} = \frac{\boldsymbol{P_d - Mean_d}}{\boldsymbol{SD_d}} \tag{4}$$

Where, $\boldsymbol{P_d}$ indicates the specific data in the list, and $\boldsymbol{Avg_V}$ indicates the average value. After normalization, the data values are standardized within the range of 0 to 1, and this filtered dataset can be used for optimization and classification operations.

3.2 Harris Hawks Optimization (HHO)

After preprocessing, an advanced meta-heuristic optimization technique, named as, Harris Hawks Optimization (HHO) technique is used to optimally select the features for improving the performance and efficiency of classification. Typically, the HHO is one of the population based gradient –free optimization mechanism extensively used in different types of applications for solving the complex optimization problems. Moreover, this technique includes the phases of exploration and exploitation, in which the exploration is executed before exploitation. During this phase, the harris hawks can identify and detect the preys through their powerful eyes. Among other optimization techniques, it has an increased convergence rate, efficiency, and reaches the best optimal solution with reduced number of iterations. In this methodology, the hawks of harries are considered as the best solutions, where each step is associated to the appropriate prey of the candidate solution. After parameter initialization, position vector of hawks are estimated by using the following model:

$$\boldsymbol{P}\ (\boldsymbol{k}+\boldsymbol{1}) = \begin{cases} \boldsymbol{P_{Rh}(k) - r1|P_{Rb}(k) - 2r_2 C(k)|} & \boldsymbol{x < 0.5} \\ \boldsymbol{\left(P_{Rb}(k) - P_{Avp}(k)\right) - r3\left(L_B + r4(U_B - L_B)\right)} & \boldsymbol{x < 0.5} \end{cases} \tag{5}$$

Where, $\boldsymbol{P}\ (\boldsymbol{k}+\boldsymbol{1})$ indicates the position of hawks in the next iteration k, $\boldsymbol{P_{Rb}(k)}$ denotes the position of rabbit, $\boldsymbol{r1, r2, r3, r4, x}$ are the random numbers (0 to 1), $\boldsymbol{P_{Avp}}$ represents the average position, $\boldsymbol{P_{Rh}}$ is the randomly selected hawk in the current population, $\boldsymbol{U_B}$and $\boldsymbol{L_B}$ are the upper and lower bounds respectively. Then, the rabbit motions is simulated according to the random values in each iteration. After that, the arbitrary location is created, and the average distance value is computed with respect to different factors. Moreover, the proposed length of momentum was estimated based on the lower bound in the rule. Moreover, the elements in the random scaling coefficient are considered for formulating the additional patters with the specified regions. Consequently, the hawks reach the final average position as shown in below:

$$\boldsymbol{P_{Avp}(k)} = \frac{\boldsymbol{1}}{\boldsymbol{No_{Hw}}}\sum_{\boldsymbol{l=1}}^{\boldsymbol{No_{Hw}}} \boldsymbol{P_l(k)} \tag{6}$$

Where, $\boldsymbol{No_{Hw}}$ denotes the number of hawks, and $\boldsymbol{P_l(k)}$ is the location of each hawk at iteration k. Here, the least intrinsic rules have been used to determine the median position. Furthermore, this model incorporates the exploration and exploitation capabilities according to the loss of energy as computed below:

$$\boldsymbol{R_E} = \boldsymbol{2R_{E0}}\left(\boldsymbol{1} - \frac{\boldsymbol{m}}{\boldsymbol{M}}\right) \tag{7}$$

Where, $\boldsymbol{R_E}$ indicates the prey's escape energy, $\boldsymbol{R_{E0}}$ is the initial energy level, and $\boldsymbol{M}$ indicates the maximum number of iterations. In this model, the following rules are applied to estimate the hawks behavior:

$$\boldsymbol{P(k+1)} = \boldsymbol{\Delta P(k) - R_E|GP_{Rb}(k) - P(k)|} \tag{8}$$

$$\boldsymbol{\Delta P(k)} = \boldsymbol{P_{Rb}(k) - P(k)} \tag{9}$$

Where, $\boldsymbol{\Delta P(k)}$ indicates the position of rabbit in current position at iteration k. Then, the current position is updated by using the following model:

$$P(k+1) = P_{Rb}(k) - R_E|\Delta P(k)| \tag{10}$$

According to the updated position, the optimal fitness value is identified from the available features. In addition to that, the dimensionality of features have been efficiently reduced by optimally choosing the characteristics, and its mutual importance is validated as follows:

$$MutS\ (G, CF) = \frac{1}{|G|}\sum_{Re_F \in CF} H(Re_F, CF) \tag{11}$$

Where, $MutS\ (G, CF)$ denotes the mutual significance, CF is the common feature, Re_F defines the relevant feature, and G is considered as the selected function with fitness value. Based on this operation, the best optimal function is identified to select the most suited features, which can be used for training the classifier.

3.3 Pivotal Decision Tree (PDT) Classification

After feature optimization, an advanced and efficient Pivotal Decision Tree (PDT) based classification methodology is implemented to detect diabetes according to the optimized feature set. The PDT is a kind of machine learning classification model mainly used for solving the complex classification problems. Typically, the machine learning techniques are extensively used in many detection/prediction application systems for solving the given problems. Also, it obtains the features as the input for processing, and produced the predicted or classified label as the output. In the conventional works, the different types of machine learning techniques are developed for diabetes prediction and classification. It includes SVM, NB, DT, RF, NB, and etc, but it limits with the key problems of increased mis-prediction results, high error rate, requires high time consumption, and minimal performance outcomes. Hence, the proposed work intends to develop a new PDT technique for accurately predicting diabetes with reduced computational and time complexity. The sample tree construction architecture is shown in Fig 3, which comprises the root node, internal node and leaf node.

Algorithm I – PDT based classification

Input: Optimized feature set F_s;
Output: Classified label L;
Procedure PDT construction
Repeat
Maximum $I_g \leftarrow 0$; //Where, I_g – Information gain
$S_I(X) \leftarrow Null$; //Where, S_I – Split data
$c \leftarrow En_t(Set)$; //Where, En_t – Entropy
For all attributes k in Set
Estimate gain as, $G \leftarrow I_g(k, c)$
If $(G > maximum\ I_g)$ then
Maximum $I_g \leftarrow G$
$S_I(X) \leftarrow k$
End if;
End for;
Partition $(Set, S_I(X))$
Repeat all partitions;
End procedure

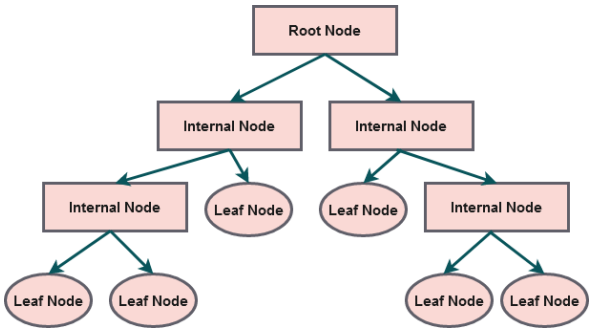

Fig 3. PDT construction architecture model

4. Results and Discussion

This section validates the performance and results of the proposed IHHO-PDT technique using various evaluation indicators such as accuracy, precision, recall, sensitivity, f1-score and time. Also, the obtained results are compared with the recent state-of-the-art model approaches for proving the betterment of the proposed technique. For this analysis, the PIMA Indian dataset is considered, which comprises various health features of the diabetes patients as shown in Table 1. Also, the confusion matrix for conventional and proposed classification techniques are depicted in Fig 3 with respect to the true label and predicted label. According to this analysis, it is observed that the proposed PDT technique outperforms the other classifiers with accurate prediction results.

Table 1. PIMA Indian diabetes dataset

Attribute No.	Type of risk	Range (Min to Max)
1	Number of time pregnant	0 to 17
2	Plasma glucose concentration	44 to 199
3	Diastolic blood pressure (mm Hg)	24 to 122
4	Triceps skinfold thickness (mm)	7 to 99
5	2-Hours Serum Insulin (mu U/mL)	14 to 846
6	Body Mass Index (BMI)	18.2 to 67.1
7	Diabetes Pedigree Function	0.07 to 2.42
8	Age (Years)	21 to 81
9	Class	1 – Positive, 0 – Negative

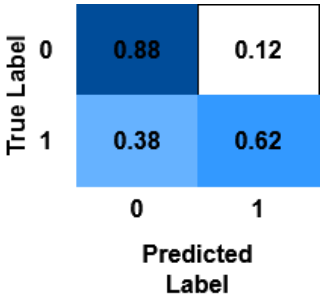

Fig 4 (a). Confusion matrix for SVM

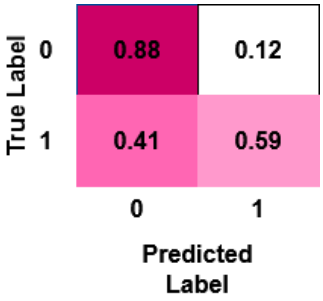

Fig 4 (b). Confusion matrix for RF

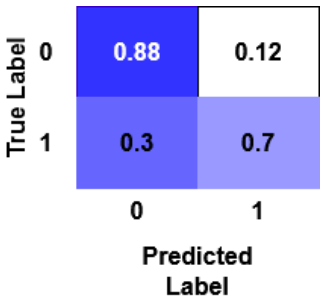

Fig 4 (c). Confusion matrix for LDA

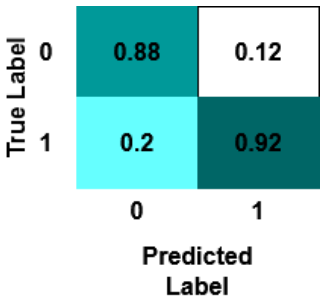

Fig 4 (d). Confusion matrix for PDT

Table 2 and Fig 5 compares the accuracy of optimization integrated classification approaches. Typically, the accuracy is one of the most essential parameter used to determine the performance of classifier. The overall performance of the detection system is determined according to the increased value of accuracy, which is estimated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (12)$$

Where, TP – True Positives, TN – True Negatives, FP – False Positives, and FN – False Negatives. From the analysis, it is evident that the proposed IHHO-PDT technique provides an increased accuracy over the other techniques. Due to the proper preprocessing and optimization processes, the accuracy of classifier is highly increased in the proposed detection system.

Table 2. Accuracy analysis of optimization integrated classifiers

Optimization based Classifiers	Accuracy (%)
Hybrid cuckoo-firefly	81
Feed forward NN	82
NB	79.56
LDA, MWSVM	89.74
GA-NN	87.76
K-means + DT	90.03
K-Means + PCA	72
Proposed	98

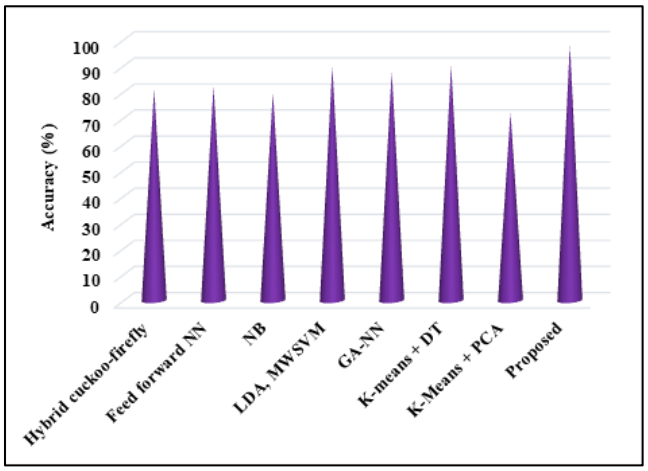

Fig 5. Accuracy of optimization based classifiers

Table 3 and Fig 6 presents the overall comparative analysis of conventional [35] and proposed classification techniques, which includes the parameters of sensitivity, specificity, accuracy, precision, recall, f1-score, and time. The measures are estimated as follows:

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (13)$$

$$Specificity = \frac{TN}{TN+FP} \times 100\% \quad (14)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (15)$$

$$F1 - score = \frac{2 \times Precision \times Sensitivity}{Precision+Sensitivity} \times 100\% \quad (16)$$

The above mentioned measures are extensively used in many detection systems for evaluating the performance and efficiency of classifier. Moreover, the overall better system performance of the diabetes detection is determined according to the improved values of these measures. Based on the estimated results, it is observed that the proposed IHHO-PDT technique overwhelms the other machine learning classifiers with increased performance results. Due to the inclusion of an efficient preprocessing and optimization techniques, the performance of the proposed diabetes detection system is highly improved in this work.

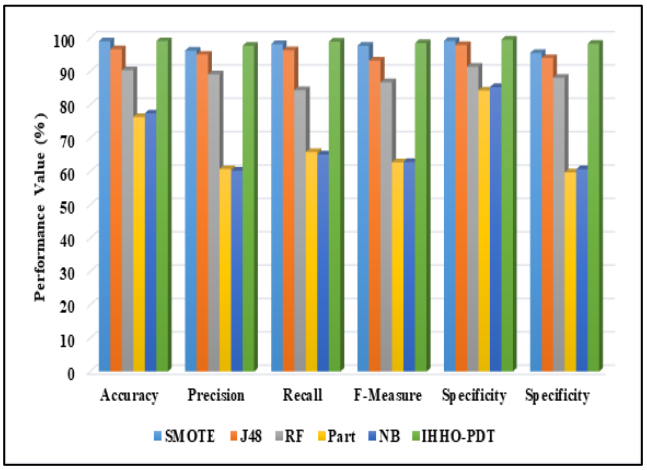

Fig 6. Overall comparative analysis

Table 3. Overall coparative analysis

Measures	SMOTE	J48	RF	Part	NB	IHHO-PDT
Accuracy	99.07	96.62	90.34	76.33	77.43	99.08
Precision	96.23	95.06	89.09	60.72	60.23	97.65
Recall	98.24	96.35	84.43	65.81	65.09	98.9
F-Measure	97.71	93.21	86.78	62.76	62.88	98.5
Specificity	99.14	97.86	91.43	84.29	85.28	99.5
Specificity	95.52	94.03	88.06	59.70	60.65	98.3
Time (ms)	0.1	0.2	3.66	0.25	0.5	0.09

Fig 7 validates the execution time of conventional and proposed machine learning classifiers. The time is also considered as one of the most essential parameter used to determine the performance of prediction system. Also, it can be estimated by the total amount of time required to accomplish the system operations. From the analysis, it is evident that the time required for the proposed diabetes detection system is efficiently reduced, when compared to the other techniques. Fig 8 and Table 4 validates the detection accuracy of existing [23] and proposed machine learning technique. Moreover, the training and testing accuracy measures are validated for both conventional and proposed machine learning techniques as shown in Fig 9 and Table 5. Normally, the classifier training is performed with the input of features obtained from the dataset. In the proposed framework, an optimized feature set is generated from the preprocessed dataset, which helps to increase the training and testing accuracy of classifier. Hence, the proposed IHHO-PDT technique provides an increased accuracy for both training and testing operations, when compared to the other machine learning approaches.

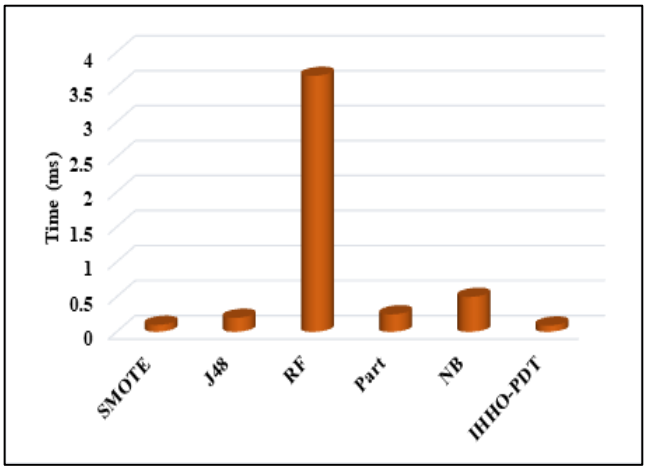

Fig 7. Time analysis

Table 4. Detection accuracy

Methods	Detection Accuracy (%)
Decision Tree (DT)	86
Gaussian Naïve Bayes (GNB)	93
Linear Discriminant Analysis (LDA)	94
Support Vector Classifier (SVC)	60
Random Forest (RF)	91
Extra Trees (ET)	91
Ada Boost (AB) Classifier	93
Perceptron Classifier (PC)	76
Logistic Regression (LR)	96
Gradient Boost (GB) Classifier	93
Bagging	90
K-Nearest Neighbor (KNN)	90

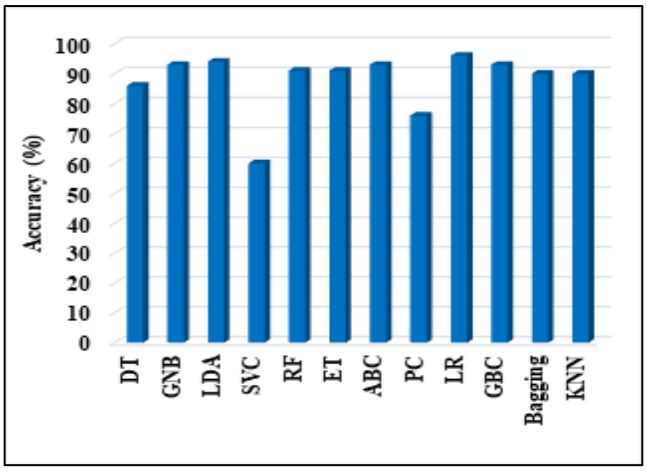

Fig 8. Accuracy analysis

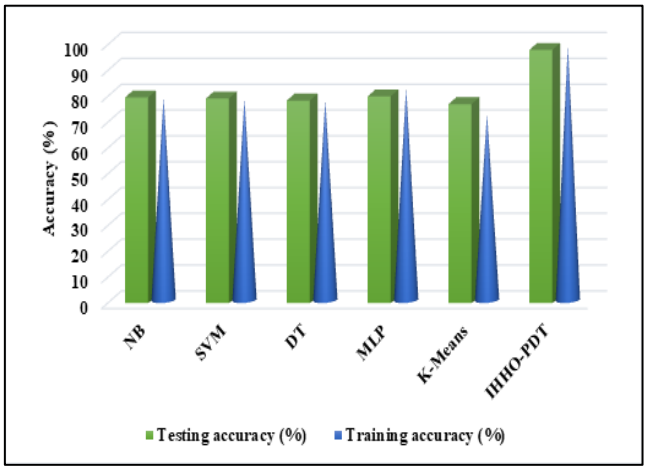

Fig 9. Training and testing accuracy

Table 5. Training and testing accuracy

Classifiers	Testing Accuracy (%)	Training Accuracy (%)
NB	79.6	78.6
SVM	79.2	78
DT	78.4	77.2
MLP	80	82
K-Means	77	72
IHHO-PDT	99	98.5

5. Conclusion

This paper presents a new prediction framework for detecting diabetes from the PIMA Indian dataset. The novel contribution of this work is to implement an advanced and efficient data mining techniques for developing the diabetes prediction system. The proposed framework includes the stages of preprocessing, feature optimization, and classification. Here, an ICN technique is deployed at first for preprocessing the given dataset to eliminate the irrelevant attributes and to normalize the contents. Based on this process, the overall quality of data has been improved, which can be further used for optimization and classification operations. During optimization, an IHHO technique is used to optimally select the features for training the classifier. Here, the purpose of implementing this approach is to reduce the computational complexity and, time required for developing the prediction system. Furthermore, an efficient machine learning based classification model, named as, PDT is used for predicting the classified label as whether diabetes or non-diabetes. The primary advantages of the proposed IHHO-PDT technique are as follows: reduced complexity, minimal time consumption, optimized performance rate, and high efficiency. Moreover, the results of the proposed detection system is validated and compared by using various evaluation measures such as precision, accuracy, recall, time, f1-score, sensitivity, and specificity. Then, the recent state-of-the-art machine learning techniques are compared with the proposed model. According to the analysis, it is observed that the proposed IHHO-PDT technique outperforms the other approaches with improved results, which shows the overall efficiency and performance of the proposed work. In future, this work can be extended by implementing a deep learning based classification model for developing the diabetes prediction system.

References

[1]. A. Paleczek, D. Grochala, and A. Rydosz, "Artificial breath classification using XGBoost algorithm for diabetes detection," Sensors, vol. 21, p. 4187, 2021.

[2]. G. Kalyani, B. Janakiramaiah, A. Karuna, and L. Prasad, "Diabetic retinopathy detection and classification using capsule networks," Complex & Intelligent Systems, pp. 1-14, 2021.

[3]. S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40-46, 2021.

[4]. K. M. Aamir, L. Sarfraz, M. Ramzan, M. Bilal, J. Shafi, and M. Attique, "A Fuzzy Rule-Based System for Classification of Diabetes," Sensors, vol. 21, p. 8095, 2021.

[5]. D. Rahhal, R. Alhamouri, I. Albataineh, and R. Duwairi, "Detection and Classification of Diabetic Retinopathy Using Artificial Intelligence Algorithms," in 2022 13th International Conference on Information and Communication Systems (ICICS), 2022, pp. 15-21.

[6]. A. Sharma, K. Guleria, and N. Goyal, "Prediction of diabetes disease using machine learning model," in International Conference on Communication, Computing and Electronics Systems, 2021, pp. 683-692.

[7]. S. Shafi and G. A. Ansari, "Early prediction of diabetes disease & classification of algorithms using machine learning approach," in Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), 2021.

[8]. Kumari, S. S. ., and K. S. . Rani. “Big Data Classification of Ultrasound Doppler Scan Images Using a Decision Tree Classifier Based on Maximally Stable Region Feature Points”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 76-87, doi:10.17762/ijritcc.v10i8.5679.

[9]. D. K. Choubey, S. Tripathi, P. Kumar, V. Shukla, and V. K. Dhandhania, "Classification of Diabetes by Kernel based SVM with PSO," Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), vol. 14, pp. 1242-1255, 2021.

[10]. R. Saxena, S. K. Sharma, M. Gupta, and G. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[11]. P. Tiwari and V. Singh, "Diabetes disease prediction using significant attribute selection and classification approach," in Journal of Physics: Conference Series, 2021, p. 012013.

[12]. K. Thaiyalnayaki, "Classification of diabetes using deep learning and svm techniques," International Journal of Current Research and Review, vol. 13, p. 146, 2021.

[13]. H. N. K. Al-Behadili and K. R. Ku-Mahamud, "Fuzzy unordered rule using greedy hill climbing feature selection method: An application to diabetes classification," Journal of Information and Communication Technology, vol. 20, pp. 391-422, 2021.

[14]. A. Bilal, G. Sun, Y. Li, S. Mazhar, and A. Q. Khan, "Diabetic retinopathy detection and classification using mixed models for a disease grading database," IEEE Access, vol. 9, pp. 23544-23553, 2021.

[15]. A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," Computers in Biology and Medicine, vol. 136, p. 104664, 2021.

[16]. X. Wang, M. Zhai, Z. Ren, H. Ren, M. Li, D. Quan, et al., "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," BMC medical informatics and decision making, vol. 21, pp. 1-14, 2021.

[17]. J. B. Awotunde, F. E. Ayo, R. G. Jimoh, R. O. Ogundokun, O. E. Matiluko, I. D. Oladipo, et al., "Prediction and classification of diabetes mellitus using genomic data," in Intelligent IoT systems in personalized health care, ed: Elsevier, 2021, pp. 235-292.

[18]. S. Singh, B. Rathore, H. Das, S. Agrawal, D. Bhutia, V. Maan, et al., "An Improved Convolutional Neural Network for Classification of Type-2 Diabetes Mellitus," in Proceedings of the Third International Conference on Information Management and Machine Intelligence, 2023, pp. 417-424.

[19]. J. Chaki, S. T. Ganesh, S. Cidham, and S. A. Theertan, "Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review," Journal of King Saud University-Computer and Information Sciences, 2020.

[20]. R. Kamalraj, S. Neelakandan, M. R. Kumar, V. C. S. Rao, R. Anand, and H. Singh, "Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm," Measurement, vol. 183, p. 109804, 2021.

[21]. T. Chauhan, S. Rawat, S. Malik, and P. Singh, "Supervised and unsupervised machine learning based review on diabetes care," in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 581-585.

[22]. Patil, V. N., & Ingle, D. R. (2022). A Novel Approach for ABO Blood Group Prediction using Fingerprint through Optimized Convolutional Neural Network. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 60–68. https://doi.org/10.18201/ijisae.2022.268

[23]. S. S. Reddy, N. Sethi, R. Rajender, and V. Vetukuri, "Non-invasive Diagnosis of Diabetes Using Chaotic Features and Genetic Learning," in International Conference on Image Processing and Capsule Networks, 2022, pp. 161-170.

[24]. H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," Clinical eHealth, vol. 4, pp. 12-23, 2021.

[25]. N. A. Libre. (2021). A Discussion Platform for Enhancing Students Interaction in the Online Education. Journal of Online Engineering Education, 12(2), 07–12. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/49

[26]. A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," Procedia Computer Science, vol. 165, pp. 292-299, 2019.

[27]. T. N. Joshi and P. Chawan, "Diabetes prediction using machine learning techniques," Ijera, vol. 8, pp. 9-13, 2018.

[28]. M. Alehegn, R. Joshi, and P. Mulay, "Analysis and prediction of diabetes mellitus using machine learning algorithm," International Journal of Pure and Applied Mathematics, vol. 118, pp. 871-878, 2018.

[29]. Agarwal, D. A. . (2022). Advancing Privacy and Security of Internet of Things to Find Integrated Solutions. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 05–08. https://doi.org/10.17762/ijfrcsce.v8i2.2067

[30]. H. R. Abdulqadir, A. M. Abdulazeez, and D. A. Zebari, "Data mining classification techniques for diabetes prediction," Qubahan Academic Journal, vol. 1, pp. 125-133, 2021.

[31]. G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, "Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by cross-validation methods," Machines, vol. 7, p. 74, 2019.

[32]. K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," Materials Today: Proceedings, 2021.

[33]. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Frontiers in genetics, vol. 9, p. 515, 2018.
[34]. M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," Computer Methods and Programs in Biomedicine, vol. 202, p. 105968, 2021.
[35]. J. Ramesh, R. Aburukba, and A. Sagahyroon, "A remote healthcare monitoring framework for diabetes prediction using machine learning," Healthcare Technology Letters, vol. 8, pp. 45-57, 2021.
[36]. J. J. Boutilier, T. C. Chan, M. Ranjan, and S. Deo, "Risk stratification for early detection of diabetes and hypertension in resource-limited settings: machine learning analysis," Journal of medical Internet research, vol. 23, p. e20123, 2021.
[37]. A. A. Abaker and F. A. Saeed, "A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications," Informatica, vol. 45, 2021.
[38]. M. H. Mahmoud, S. Alamery, H. Fouad, A. Altinawi, and A. E. Youssef, "An automatic detection system of diabetic retinopathy using a hybrid inductive machine learning algorithm," Personal and Ubiquitous Computing, pp. 1-15, 2021.
[39]. H. Naz and S. Ahuja, "SMOTE-SMO-based expert system for type II diabetes detection using PIMA dataset," International Journal of Diabetes in Developing Countries, pp. 1-9, 2021.