Yugoslav Journal of Operations Research
35 (2025), Number 4, 895-918
DOI: https://doi.org/10.2298/YJOR240615001A

Research Article

PREDICTION OF TYPE 2 DIABETES USING SUPPORT VECTOR MACHINE WITH ENHANCED LEVY FLIGHT BASED FRUITFLY OPTIMIZATION ALGORITHM AND FEATURE SELECTION APPROACHES

Ashok KUMAR M*
Computer Science and Software Engineering, Skyline University Nigeria, Kano City, Kano State, Nigeria, ashok.kumar@sun.edu.ng; williamashok@gmail.com, ORCID: 0000-0001-6869-4975

Vijay Arputharaj J
Department of Computer Science, CHRIST (Deemed to be University), Bangalore, India phdvij@gmail.com, ORCID: 0000-0002-3842-5835

Sathya V
Department of Computer Science, Navarasam college of Arts and Science, Erode, Tamilnadu India, vbs.sathya@gmail.com, ORCID: 0009-0006-1724-5647

Dalvin Vinoth KUMAR
Department of Statistics and Data Science, Christ University, Bangalore, India, dalvin.vinoth@christuniversity.in, ORCID: 0000-0003-0768-3097

Shanmugam SUNDARARAJAN
Department of Economics and Entrepreneurship, Skyline University Nigeria, Kano City, Kano State, Nigeria, s.sundararajan@sun.edu.ng, ORCID: 0000-0001-9712-6635

Sivanantham V
*Department of Computer Science, Periyar University, Salem, India,vmpsiva@gmail.com,*ORCID: 0009-0003-1985-1163

Sanjoy KUMAR PAL
Department of Biological Sciences SSIT, Skyline University Nigeria Kano City, Kano State, Nigeria, sanjoypal@yahoo.com, ORCID: 0000-0003-1436-6880

Received: June 2024 / Accepted: October 2024

* Corresponding author

Abstract: Researchers have been leveraging various data analytics methods for Diabetes mellitus (DM) diagnosis, prognosis and management. The data analytics paradigm has become advanced and automated with the emergence of machine learning (ML) and deep learning (DL) algorithms. With new techniques, the prediction accuracy of ML models for various real-world problems has increased significantly. In our previous work, we introduced and investigated the Improved K-Means with Adaptive Divergence Weight Binary Bat Algorithm to create an innovative diagnosis system. Across several problem scenarios, the performance of this algorithm is much better in terms of speed. However, this algorithm's accuracy of data categorization comes below expectations. To achieve high classification accuracy, the objective of this study work is to concentrate on methods and strategies. This aim is fulfilled through a Support Vector Machine (SVM) with an Enhanced Levy Flight-based Fruitfly Optimization model. This novel model improves diabetes prediction accuracy and can be applied to regressions, classifications, and other tasks. The nearest training data points' distances should be greater as this can lower classifiers' generalization errors. Missing values in datasets are retrieved using the Adaptive Neuro Fuzzy Inference System (ANFIS). A new algorithm called the Enhanced Inertia Weight Binary Bat Algorithm (EIWBBA) is introduced to optimize feature spaces and eliminate unimportant aspects. Further on, a novel feature selection technique is introduced by using the Enhanced Generalized Lambda Distribution Independent Component Analysis (EGLD-ICA). The classification uses a Support Vector Machine with an Enhanced Levy flight-based Fruitfly Optimization Algorithm (SVM-ELFFOA). The SVM-ELFFOA classification techniques are implemented using MATLAB software. It is evident that the discussed IKM-EIWBBA+SVM-ELFFOA classifier produces much better values of the accuracy of 93.50%, while the available IKM-EIWBBA+SVM yields 91.87%, IKM-ADWFA+LR renders 90.50%, and IKM+LR renders just 85.00%. From the simulation experiment, the proposed classification techniques implemented in MATLAB software and according to comparative data, this suggested model has a higher prediction accuracy of 93.50% compared to existing classification methods.

Keywords: K-means algorithm, enhanced inertia weight binary bat algorithm, adaptive neuro-fuzzy inference system, diabetes mellitus prediction, enhanced generalized lambda distribution independent component analysis.

MSC: 68T05, 68T20, 62H30, 62H30.

1. INTRODUCTION

DM, popularly known as Diabetes, is one of the diseases that affect a huge population in the world. 2017 Diabetes [1] has affected over 425 million people, which is quite big. Diabetes and its respective severity have proved to be fatal reaching about 4 million individuals in a single year. In India, Diabetes has affected 74 million people, and India is called the "Diabetes Capital of the World". If proper remediation is not taken proactively and pre-emptively, therefore the number of individuals who would be negatively impacted by Diabetes may rise to over 629 million people around the world in the year 2045 as per prediction.

Earlier, disease occurrences were predicted and classified using ML algorithms [2]. To produce competent prediction models, feature engineering should be carried out first on the input data to acquire robust features [3]. There were a few important issues that needed to be resolved here. The primary objective of this research project is to use the

SVM-ELFFOA technique to forecast diabetes onset with improved accuracy. This approach also helps in predicting the risk factors and the severity of diabetics. Therefore, the proposed research concentrates on SVM-ELFFOA-based classifications for DM classifications and predictions [4]. This method has the wherewithal to eliminate the issues of existing classifiers.

To facilitate early detection and successful intervention, the creation of accurate and dependable prediction models is essential in light of the rising incidence of Type 2 diabetes [5]. However, existing methods often encounter major obstacles, such as poor feature selection, optimization techniques that converge too soon, and insufficient scalability when dealing with massive datasets. Support Vector Machines (SVMs) and other traditional machine learning models are susceptible to diminished accuracy due to the addition of irrelevant or necessary features [6]. In addition, popular optimization methods like Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) sometimes converge too soon, which means they don't have a chance to investigate all possible solutions and end up with less accurate models [7]. This research highlights these problems and suggests a new solution: combining SVM with sophisticated feature selection techniques and an Enhanced Levy Flight-based Fruitfly Optimization Algorithm (ELFFOA) [8]. This approach seeks to address the drawbacks of increasing the efficacy and precision of Type 2 diabetes prediction current models by strengthening the optimization process's exploration capabilities and selecting just the most important features [9].

The paper's primary innovation and application is

- Creating a Type 2 Diabetes Prediction and Classification using an SVM-ELFFOA)
- Introducing ANFIS for retrieving the missing values in datasets and EIWBBA to optimize feature spaces and eliminate unimportant aspects.
- The experimental results are performed, and the suggested SVM-ELFFOA model increases the precision, recall, F1-score and accuracy compared to existing models.

It is the structure of the remaining research: Section 1 explains a detailed summary of diabetes mellitus and its causes, the importance of prediction and the analysis of classification techniques. Section 2 overviews the relevant technical literature and its classification approaches. Section 3 explains the proposed prediction of DM using a Support Vector Machine with Enhanced Levy Flight-based Fruitfly Optimization. A description of the findings is provided in Section 4. Conclusion and future work are summed up in Section 5.

2. LITERATURE REVIEW

A large range of models based on learning capacity, adaptability, complexity, and scalability may be found in ML domains. Many well-known machine learning (ML) algorithms and related ensembles, including stochastic gradient, Random Forest (RF), decision trees (DT), regression, Support Vector Regressors (SVR), and others, may be used to address a wide range of real-world issues. Hybrids of these methods have frequently been created to address their inherent shortcomings and offer robustness, resilience, and adaptability.

In this section, some of the most recent methods for predicting diabetes mellitus are presented. Various techniques have been implemented to improve the DM diagnostic system's accuracy. The classifier's architecture is changed, and parameters are manipulated using the most available techniques to enhance performance. The content below briefly outlines a few important techniques in diabetic mellitus prediction.

Sisodia et al. [10] suggested using ML classification methods, including Naive Bayes (NB), SVM, and DT, to identify diabetes early in the experimental process. The experiment was conducted on a Pima Indians Diabetes Database (PIDD) created from the UCI ML database.

Alehegn et al. [11] proposed using three predictive-based algorithms, namely SVM, Naive Net, and Decision Stumps, in their Proposed Ensemble Method (PEM) created by combining many distinct methodologies and strategies to create an ensemble hybrid model. A total of 768 records available online and acquired from Pima Indian Diabetes DataSet (PIDD) were utilized, where PEM achieved an accuracy of 90.36%.

Established classification algorithms such as Naïve Bayes, One R and Zero R are sufficient for predicting diabetic mellitus by Mounika et al. [12]. As most parameters do not deal with definite categorical values, it is found that the linear regression model is sufficient. In the Weka tool, various classification algorithms are introduced to approximate the efficiency and accuracy of each algorithm.

Pavate et al. [13] have introduced an efficient technique for forecasting diabetes and more risk-level complications. A few approaches have been used in this system to produce a robust prediction system.

The best characteristics subset was selected using genetic algorithm variations based on the nearest neighbour groupings despite the proposed system's bias, and the results were quite satisfactory. The positive outcomes showed precision (86.95%), accuracy (95.50%), and sensitivity (95.83%) on data that validated the system's ability to anticipate the illness.

By integrating two methods, the Bayesian classification and the Multi-layer Perceptron, for diagnosing diabetes diabetes-mellitus, Kumar Dewangan et al. [14] introduced an ensemble model. MLP is a basic perceptron creation in which extra hidden layers (in addition to the input and output layers) are applied.

Class membership probabilities, such as the chance that a certain tuple belongs to a particular class, may be predicted using statistical classifiers like the Bayesian Net. With six characteristics from the simulation data, this model obtains the highest accuracy of 81.89%, the best sensitivity of 64.10%, and the highest specificity of 90.90%.

Saravananathan et al. [15] recommended techniques for classification (for estimating the severity of diabetes): k-nearest Neighbor (kNN), J48, Support Vector Machines (SVM), Classification and Regression Tree CART. Classification methods use diabetic data as input and supporting elements during the predefined collection of modules to explore the layout of these classification methods. The predictive algorithms Joshi et al. [16] proposed use KNN, NB, RF, and J48. These algorithms create an ensemble hybrid framework to increase performance and precision by combining individual techniques/methods into one. The primary objective of this evaluation is the prediction of Diabetes disease and then comparing it with a single algorithm.

Sneha et al. [17] use ML to design a prediction algorithm, and an optimal classifier is obtained for producing the closest results. Diabetes nautilus can be detected early using the proposed technique of cognitively selecting attributes.

Around 98.00% specificity value is produced by RF, and 98.20% is produced using the DT algorithm, which gives the best results for diabetic data analysis. A better accuracy of 82.30% is produced by Naïve Bayesian (NB) outcome.

Various classification algorithms like RF, Zero R, J., MLP and Naïve Bayes are proposed by Hina et al. [18]. From a specified dataset, knowledge can be extracted for research and the generation of comprehensive and intelligent results. The multi-layer perception (MLP) function can produce highly effective results. Fewer errors are produced due to high processing time; every node's weights must be computed.

Asgarnezhad et al. [19] present an effective pre-processing technique with a popular DM data set. This includes attribute subset selection techniques and missing value replacement techniques. The applied classifier's performance can be enhanced using the proposed technique. Concerning precision and accuracy, it outperforms traditional techniques.

Cárdenas-Cabrera Jorge et al. [20] recommended cutting-edge process integration techniques for characterization and adaptation. The paper presents fresh methods for controlling integrated systems and characterizing processes. In the business world, integrative systems are commonplace. These include temperature, pressure, concentration, and level control systems. Despite the abundant literature on the topic, no universally accepted standard has been developed, leaving this issue unresolved. Applying characterization approaches for self-regulated processes and deriving the response of the integrated system make up the suggested characterization method. The λ tuning rules are used to derive the formulae for PD control modifying, which forms the basis of the suggested tuning approach. Simulation and experimental testing confirm the suggested approaches' efficacy.

Farahani et al. [21] proposed the Efficient Market Hypothesis-Based Hybrid Metaheuristic Artificial Neural Networks for Predicting Stock Prices. In order to find the best optimization indicators, the Genetic Algorithm (GA) is used. Neural network (NN) fine-tuning includes indicator selection, particle swarm optimization (PSO), and harmonic search (HS), all of which work together to optimize the number of hidden layers, weights, and the network's error threshold. To evaluate the recommended model's efficiency and select the optimal model using error criteria, the author uses eight estimate criteria for error evaluation. One novel feature of this study is that it uses the most important Iranian enterprises as a statistical population and tests market efficiency. When compared to other algorithms, the testing findings show that a hybrid ANN-HS algorithm provides the most accurate stock price predictions.

Maria Lincy Jacquline and Natarajan Sudha [22] recommended the Weighted Fuzzy C Means and Enhanced Adaptive NeuroFuzzy Inference Chronic Kidney Disease (CKD) Classification. Novel approaches to disease classification, such as the Fruit Fly Optimization Algorithm (FFOA) and the Multi-Kernel Support Vector Machine (MKSVM), have been introduced in recent studies. When evaluating a collection, FFOA is often used to identify the most valuable aspects. MKSVM sorts medical records into distinct categories by applying predefined criteria to the information. Any number of possible deviations from the expected data set will affect the classifier's precision. However, MKSVM still produces a higher rate of inaccurately labelled results. To address these issues, a min-max normalization-based pre-processing step is used to scale-normalize the input CKD data. Afterwards, key features will be chosen using Improved FFOA (IFFOA). The chosen characteristics will be grouped using Weighted Fuzzy C

Means clustering (WFCM) to eliminate or significantly decrease misclassification. Using the EANFIS method, the final step is to categorize CKD as normal or abnormal. Impressive recall, accuracy, precision, and f-measure results show that the proposed technique works.

Reza Rasinojehdehi and Seyyed Esmaeil Najafi [23] discussed the Data Envelopment Analysis (DEA) and SVM for Advancing Risk Assessment in Renewable Power Plant Construction. Phase one of DEA involves measuring risk variables obtained from Failure Modes and Effects Analysis (FMEA). Better discrimination capabilities for decision units result from this method's ability to circumvent certain FMEA shortcomings while removing some DEA constraints. Finally, a SVM is trained to keep an eye on things, and the whole thing comes to a close with risk mitigation and oversight procedures devised with Iran's solar energy scenery in mind.

Mojtaba et al. [24] discussed how to choose the Open Innovation Method in the Automotive Industry using the Adaptive Network-Based Fuzzy Inference System. The technical knowledge level of the firm, the complexity of the part's technology, and nine potential outputs are the two inputs that fuzzy reasoning considers, including various open innovation methods. This enables the system to extract a technique appropriate for the requirements of the organization using the current regulations. To develop the models in this study, 50% of the data were used as training data, while 50% were utilized as test data. There was a 90% success rate in using the created model to choose open innovation strategies. Thus, the model shown is suitable for selecting an open innovation strategy for the automotive industry.

Sahabul Alam [25] presented the Trusted Fuzzy Routing Scheme in Flying Ad-hoc Network (FANETs). Due to the rising demand for dynamic and flexible FANET communications, a trustworthy and bioinspired transmission approach has been created. The fitness theory measures direct trust. Indirect trust is measured by action and believability, though. Assessing UAV behaviour is critical. For trustworthy route computation, it proposes fuzzy logic, a popular method. Fuzzy logic can categorize nodes using numerous criteria to handle complex circumstances. This approach predicts intermediate UAV locations using 3D estimations from geocaching and unicasting. This approach improves FANET performance by ensuring robustness, reliability, and path lifetime. Routers must support two FANET enhancements that decrease route lifespan. First, collaboration requires energy-intensive communication and coordination between flying nodes. Second, link disconnection from dispersion may be caused by the very dynamic 3D mobility pattern of the flying nodes. It selects trustworthy leader drones and routes leaders safely using ant colony optimization. For FANET trust management, the author offers fuzzy-based UAV behavior analytics. Simulations show that FANET delay routing overhead is lower than that of other protocols.

Ahmed A. El-Douh et al. [26] examined the Machine Learning and Association Rules under a Neutrosophic Environment for Heart Disease Prediction. To forecast the occurrence of cardiovascular illness, this research used association rules, machine learning models, and the neutrosophic analytical hierarchy process (AHP) for feature selection. The neutrosophic AHP approach determines feature weights and chooses the best characteristics. It may apply the rules between values in any dataset with the help of the association rules. Next, to choose the most effective feature to feed into machine learning models, we selected features using the neutrosophic AHP. Heart disease was predicted using nine different machine-learning models. We found that the methods with

the highest accuracy rates were random forest (RF) and decision tree (DT), with 100% accuracy. With an accuracy of 99%, 98%, and 97%, respectively, bagging, k-nearest neighbors (KNN), and gradient boosting followed. The least accurate method was SVM at 68%, followed by logistic regression and Naïve Bayes at 84% and 89%, respectively, for AdaBoosting.

Nariman Khalil et al. [27] investigated the Machine Learning Models for Prediction of Chronic Kidney Disease. Improved CKD screening and prediction might lead to better patient outcomes, lower complication rates, and a more efficient healthcare system. Still, concerns about data accessibility, integration, and ethics must be addressed before CKD prediction algorithms can be used responsibly. The use of ML techniques has been very beneficial to the medical field, particularly in the area of disease prediction. This research presents an approach for forecasting the onset of chronic kidney disease (CKD) that takes use of ML techniques. In this process, many ML models are trained and compared with respect to a number of parameters. The author used five machine learning algorithms for this: logistic regression (LR), DT, RF, SVM, and KNN. The two most accurate models, LR and KNN, both achieve 99% accuracy.

Ahmed M. Ali and Said Broumi [28] recommended Thyroid Disease Prediction and Analysis Using Machine Learning with a Multi-Criteria Decision-Making Model. This research used a decision-making framework for thyroid analysis and prediction incorporating ML and MCDM. The early detection and treatment of thyroid illness would benefit individuals all around the globe since it affects a large population. In this work, we combine ML algorithms with the MCDM technique. LR, SVM, and RF are the three ML methods used in this research. These algorithms aim to analyze and forecast cases of thyroid illness. The findings demonstrate that the RF has the utmost precision, accuracy, and F1 score. For the RF, the accuracy is 0.95. The recall score of the SVM is 1.0. The best ML algorithm is selected and ranked using several elements, including the MCDM approach. The ML algorithms are evaluated using the MCDM technique TOPSIS. The criterion weights are computed using the mean technique. Based on the MCDM technique, this study's top three ML algorithms are RF, SVM, and LR, in that order.

Samia Mandour et al. [29] recommended the Mantis Search Algorithm Integrated with Opposition-Based Learning and Simulated Annealing for Feature Selection. This paper introduces OBMSASA, a novel feature selection technique that enhances the exploration and exploitation operators of the newly disclosed mantis search algorithm by combining it with the opposition-based learning (OBL) method and simulated annealing (SA). By enhancing the exploration operator, the algorithm remains from being stuck in local minima through the OBL approach; simultaneously, the SA is used as a local search to fortify further the exploration operator, which speeds up convergence. The accuracy of the selected feature is determined using the K-nearest neighbor approach. Several performance metrics evaluate the suggested algorithm, such as the convergence curve, average fitness, computational cost, length of selected features, and standard deviation. It is compared to multiple competing optimizers and tested on 21 common datasets.

Table 1 shows the analysis of various existing feature selection and classification approaches.

The suggested study offers a unique feature selection and classification strategy that reduces unnecessary characteristics and improves classification accuracy. This helps patients anticipate diabetes early and lowers the fatality rate associated with the disease.

Table 1: Comparison of the existing Works

Author	Methods	Results	Demerits
Alehegn et al. (2018)	Well-known prediction algorithms, including SVM, Naïve Net, Decision Stump, and PEM, are used in the suggested system.	Provides 90.36 % accuracy	When more sound in the data set and the target classes overlap, it performs poorly.
Pavate and Ansari (2015)	A fuzzy rule-based system, closest neighbour algorithm, and genetic algorithm have all been utilized to design a precise prognostic system to prepare for the potential for diabetes.	The satisfactory results confirmed the effectiveness of the system in disease prediction, showing 95.83% sensitivity, 95.50% accuracy, and 86.95% specificity on opaque data.	The quality affects the data's accuracy. Large data sets may cause the prediction stage to go slowly.
Saravananathan and Velmurugan (2016)	This research uses J48, SVM, Classification and Regression Trees (CART), and kNN for diabetes data.	The J48 technique's accuracy is 67.16%, according to its results.	When more sound in the data set and the target classes overlap, it performs poorly. The quality affects the data's accuracy. Large data sets may cause the prediction stage to go slowly.
Sneha and Gangil (2019)	Identify the best classifier and develop a machine learning prediction strategy	Considering the results	It also takes a long time to train because it uses many DTs to identify the class.
Asgarnezhad et al. (2017)	An effective pre-processing approach that combines methods for selecting attribute subsets and replacing missing values	Best results with an accuracy of 84.35% and precision of 83.33%	Produce false noise edges.

3. PROPOSED METHODOLOGY

This work aims to identify and evaluate DM classifications using SVM-ELFFOA. This research focuses on reducing the severity of diabetes disease with the help of prior predictions and by enhancing the treatment process of diabetic patients. In this research work, Predictions are made using the Pima Indians Diabetes data set. To determine if a patient has diabetes or not is the objective. The decision is being facilitated through a set of particular diagnostic measurements. The primary drawback of PIMA is its

incompleteness. Neglecting any significant (decision-enabling) features can result in poor classification and, hence, poor performance. Another issue is its unpreparedness for the classification tasks. To surmount these challenges, this research gives the solution sequence as follows: the first solution tries to impute the best values in places of absent values in the data set. The second solution helps in data reduction. Figure 1 illustrates a novel system for diabetes prediction.

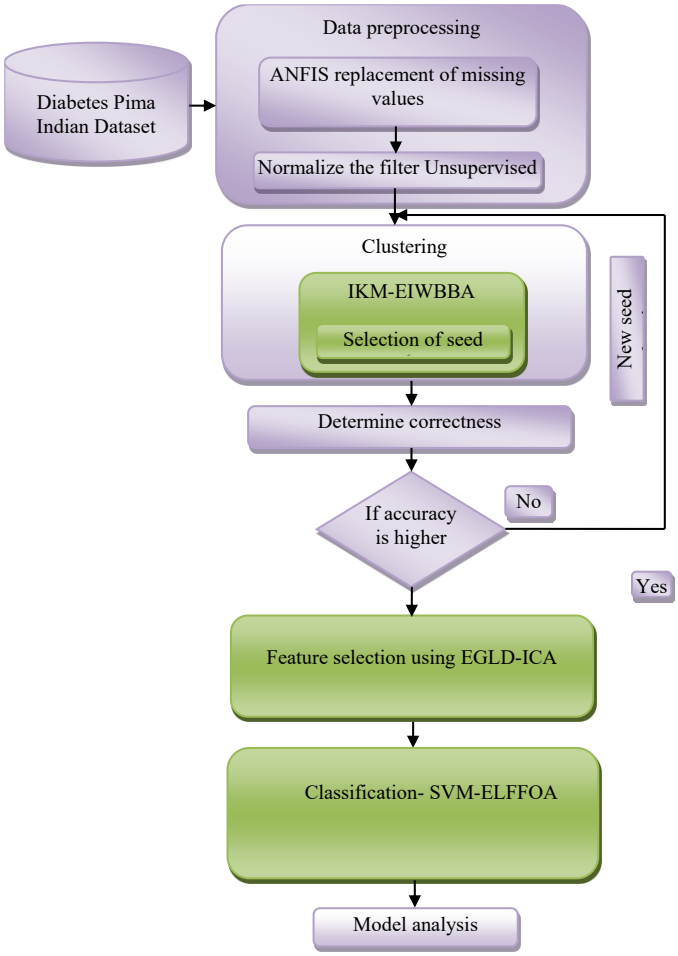

Figure 1: The architecture of the proposed system to predict diabetes

The proposed technique comprises four phases: pre-processing, categorization, feature selections, and data clustering. Reliable and proper pattern diagnosis largely depends on examining missing values in the data pre-processing stage. If the missing numbers are not properly handled at the outset, it might result in an inaccurate assessment of the patients' true category. ANFIS model is utilized to get missing values, and subsequently, EIWBBA is utilized to select seeds efficiently in an Improved K-Means algorithm. If the dimensionality reduction of the dataset is not carried out by

choosing significant features, the resulting model will become extremely complicated. Therefore, the Enhanced Generalized Lambda Distribution based Independent Component Analysis (EGLD-ICA) is proposed for feature selection. Finally, the classification activity was carried out using SVM-ELFFOA.

3.1. Data Collection

In this research work, the PIDD dataset was acquired from the UCI ML database. Several examples are gathered from the populace in the Phoenix, Arizona, USA, locations to construct this database. Information about 768 medical records is available in this PIDD. Of those, 268 records have a positive test, while the remaining records have a negative test [30]. Diabetes patients are indicated using positive results, and non-diabetic patients are indicated using negative results. Eight attributes are included in every instance and have numeric data types. The following attributes are included in this dataset: Class variable (Class), Age (Age), Diabetes pedigree function (Pedi), Body mass index (BMI), 2-h serum insulin (Insu), Triceps skinfold thickness (Skin), Diastolic blood pressure (Pres), Plasma glucose concentration at 2 h in an oral glucose tolerance test (Plas) and Number of times pregnant (Preg).

3.2. Data Pre-processing

Providing the data for machine learning (ML) is the initial stage in the process. This stage involves removing the incorrectly categorized data. To avoid the creation of incorrect or undesired rules or patterns, the data is cleaned and filtered in this instance, and the machine learning algorithm is used. An attribute for selecting a collection of characteristics with strong prediction potential is initially chosen in the pre-processing stage. All missing values are handled, and each probability is examined [31]. A suitable approach should be used to impute values instead of missing values if an attribute has more than 5% missing values. The records should not be deleted in this case. In this research work, the ANFIS model replaces missing values.

- **Replacing the Missing Values in the dataset using ANFIS**

Substituting missing values with common values usually does not produce good performance results with ML. The Artificial Neural Network (ANN) finds it more challenging to recognize the characteristics [32]. Therefore, by overcoming the problems related to incomplete values, an ANFIS can resolve this issue. The advantages of ANN and fuzzy logic (FL) are combined into one framework by ANFIS. Modelling complex patterns and comprehending nonlinear interactions, provides a quicker learning curve and adaptable interpretation abilities.

- **ANFIS Architecture**

ANFIS is a hybrid model that includes fuzzy logic's exceptional knowledge representation and inference skills combined with ANN's learning capabilities, allowing them to self-adapt their membership function to achieve maximum performance. This includes if-then rules, fuzzy logic operators, and membership functions. The use of fuzzy operators, application strategy, output aggregation, defuzzification, and input fuzzification are the five key processing stages of ANFIS operation.

ANFIS combines the potential and proven techniques available in fuzzy logic and neural networks for achieving excellent quantity and quality reasoning [33]. Therefore,

any network developed through fuzzy logic exhibits a remarkable ability to train using neural networks and variables' linguistic interpretation. Both perform simultaneous information encoding and distributed architecture in a numerical model. The rule generation depends on the fuzzy inference system.

3.3. Seed Selections using Improved K-Means Clustering Algorithm with EIWBBA (IKM-EIWBBA)

K-Means is a prominent algorithm used for clustering. It is a simple distance-based approach, and distance refers to a similarity measure that decides that a short distance will incline towards the highest similarity for revealing all objects. K from N given is the initial clusters' count [34]. The distance between the individual object and cluster centre 'm' is computed. Each object is clustered to the closest cluster based on distance applying equation (1),

$$S_i^{(t)} = \left\{ \forall j, \left|\left| x_p - m_i^{(t)} \right|\right|^2 \leq \left|\left| x_p - m_j^{(t)} \right|\right|^2 \forall j, 1 \leq j \leq k \right\} \quad (1)$$

Re-compute each cluster centre to check if they are modified, applying equation (2)

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_{j.} \quad (2)$$

Repeat the steps mentioned above until the new cluster center is similar to the one that has converged and ends the algorithm.

- **EIWBBA**

The bat's-echolocation nature influences the bat algorithm. When bats search for their prey, they decrease the loudness and increase the emitted ultrasonic sound frequency. These bats' features are utilized in the BA design [35]. The BA's fundamental steps are mathematically formulated as below. In BA, each bat includes three vectors, which consist of frequency, velocity and position vectors, which get updated during time step t as equation (3), (4), and (5):

$$V_i(t+1) = V_i(t) + (X_i(t) - Gbest)F_i \quad (3)$$

$$X_i(t+1) = X_i(t) + V_i(t+1), \quad (4)$$

Where$Gbest$indicates the best position attained till now, indicates i th bat's frequency, which is updated as below:

$$F_i = F_{min} + (F_{max} - F_{min})\beta, \quad (5)$$

It lies between the [0, 1] range and refers to a random vector with uniform distribution. Considering (2) and (4), it is apparent that multiple frequencies encourage bats' exploration potential to the optimum solution.

To resolve the early convergence problem, the Binary Bat Algorithm performs the same as the actual Binary Algorithm.

To address optimization problems in binary search space, the Binary Bat Algorithm (BBA) is provided. The BBA design is similar to the genuine BA, where velocity and frequency are expressed in continuous space. Two changes are made to real BA using BBA:

- A position vector is no longer a vector having a continuous value and is a bit string.
- For binary search space, the random function presented in (5) is no longer acceptable. Rather, an ordinary function is followed.

The expression for position update for BBA is modified to equation (6) and (7),

$$x_i^k(t+1) = \begin{cases} (x_i^k(t))^{-1} rand \leq f\left(v_i^k(t+1)\right) \\ x_i^k(t) rand > f\left(v_i^k(t+1)\right) \end{cases} \quad (6)$$

where,

$$f\left(v_i^k(t)\right) = \left|\frac{2}{\pi}\arctan\left(\frac{\pi}{2}v_i^k(t)\right)\right| \quad (7)$$

and in kth dimension,$x_i^k(t)$ indicates position and $v_i^k(t)$ indicates the velocity of ith artificial bat at iteration t and $(x_i^k(t))^{-1}$ indicates complement of $x_i^k(t)$.

The operation expressed using (8) for BBA is modified to equation (8),

$$X_{new} = X_{old} \quad (8)$$

- **Inertia weight strategies**

Where Xold represents solutions selected at random from the finest available options. Hence, inertia weights are crucial to preserving balances between local and global searches. At time steps, inertia weights determine contribution ratios between previous and current velocities. Inertia weights keep track of search situations and modify weight values in response to one or more feedback parameters.

The magnitude of velocity is managed by the use of the inertia weight approach. The following is how this strategy is stated to expressed as equation (9):

$$w = w_{max} * \exp\left(-m * \left(\frac{iter}{iter_{max}}\right)^m\right), \quad (9)$$

Where iter max represents the overall iterations count, iter indicates the present iteration's count, $w$max indicates maximal inertia values and $m$refers to a constant bigger than1. The advantage of the novel EIWBBA is that it greatly distributes the solutions onto binary search space. Also, the results achieved are highly accurate.

3.4. EGLD-ICA for Feature Selection

Independent Component Analysis (ICA) is an approach that has evolved recently. Its goal is to represent non-Gaussian data linearly to maximize the statistical independence or independence of the constituent parts. It is discovered that this representation acquires the fundamental structure of signal separation and feature extraction [36]. Y_1 and Y_2, two scalar-valued random variables are used to define the independence. When knowledge about the y_1 value does not convey any information about the y_2 value, and vice versa, these variables are considered independent. It should be noted that while this is true for variables s_1, and s_2, it is not true for mixture variables x_1, x_2.

Formally, independence is specified using probability densities. Suppose $p(y_1, y_2)$ represents y_1 and y_2's Joint Probability Density Function (PDF). Also, let $p_1(y_1)$ indicate y_1's marginal PDF, i.e. PDF of y_1 if it is taken separately expressed as equation (10):

$$p_1(y_1) = \int p(y_1, y_2) dy_2 \tag{10}$$

And for y2, it is the same. So y_1 and $y_2$are defined to be independent if and only if joint PDF is factorizable in the manner below expressed as equation (11),

$$p(y_1, y_2) = p_1(y_1) p_2(y_2). \tag{11}$$

This definition projects normally for any random variables (n), where joint density has to be a product of n terms. This specification helps derive the independent random variable's most significant property. It provided two functions, h_1 and h_2, exhibits at all times to illustrated as equation (12),

$$E\{h_1(y_1) h_2(y_2)\} = E\{h_1(y_1)\} E\{h_2(y_2)\}. \tag{12}$$

Another technique for ICA estimation, taking inspiration from information theory, involves reducing mutual information. This method is described here and demonstrated to yield the same idea of obtaining most non-Gaussian directions. Between m (scalar) random variables, the mutual information I, $y_i, i = 1 \ldots m$, is defined using the differential entropy concept.

$$I(y_1, y_2, \ldots\ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(y). \tag{13}$$

Mutual information is a common dependency measure existing between random variables in equation (13). It is equivalent to the common divergence, a widely used metric for assessing independence, between the joint density *f*(*y*) and its marginal densities product. It is always non-negative and only 0 when the variables show statistical independence. As a result, mutual information considers not just covariance but also the full dependency structure of the variable.

Mutual information can be understood by interpreting entropy as code length. Terms *H*(*yi*) provides code length for the yi when encoding is done individually, and *H*(*y*) provides code length when y gets coded in the form of a random vector, implying that all the elements are encoded using the same code value. Therefore, Mutual information indicates how much code length is reduced using the entire vector's encoding rather than individual components. Generally, reasonable codes can be acquired by encoding the vector entirely. But, in case the y_i are independent, they do not give any information about one another, and one could rather encode the variables individually without any increase in the code length. One significant property of mutual information is that in the case of an invertible linear transformation $y = Wx$:

$$I(y_1, y_2, \ldots\ldots, y_n) = \sum_i H(y_i) - H(x) - \log |\det W|. \tag{14}$$

Also, equation (14) for yi of unit variance, negentropy and entropy are different just by a constant sign. Therefore, the following expression,

$$I(y_1, y_2, \ldots\ldots, y_n) = C - \sum_i J(y_i) \tag{15}$$

As shown in equation (15), where C refers to a constant without dependency on W. This implies the basic association between mutual information and negentropy.

The classifier generalization is enhanced with the application of base transformation. Environmental variations are considered to increase the intra-class variance. Nevertheless, different ICA technique implementations that represent the source distributions change where the basis vector is located, giving the features a degree of

invariance. In brief, ICA looks for an orthonormal rotation in whitened space, giving statistical independence and preference for ICA techniques. Different effective mechanisms exist, and the Enhanced Generalized Lambda Distribution-based ICA mechanism is commonly used for smaller dimensions. In objective functions, such as probability, it is a fixed-point iteration that aids in maximizing both one- and multiple-independent components.

Enhanced Generalized Lambda Distribution ICA estimates the adaptive maximum likelihood that considers the skewness of distribution into consideration while modelling the source distributions. The modelling technique correlates the theoretical measure of independence and the practical estimator. EGLD's score function is utilized as ICA's objective function, which has to be maximized. Fixed-point and natural gradient algorithms using the EGLD framework are introduced for maximization.

The inverse distribution function is used to express the Generalized Lambda Distribution (GLD).

$$F^{-1}(p) = \lambda_1 + (p\lambda_3 - (1-p)\lambda_4)/\lambda_2 \tag{16}$$

Where equation (17), 0 <= p <= 1, distribution parameters are indicated using λ1, λ2, λ3, and λ4. When λ2 / (λ3 pλ3 – 1 – λ4 (1 – p) λ4–1) => 0, GLD holds true. Four nonlinear expressions are used for defining correlation between moments α1, α2, α3, and α4 and λ1, λ2, λ3 and λ4. These expressions solutions are obtained numerically. The inverse distribution function is where the scoring function comes from as EGLD's density function does not exist in a closed form using $p = F(y)$, where EGLD's distribution function is referred to as $F(y)$. For the p-value, the solution is computed numerically for computing the score function value of observation y, and the score function $\phi(p)$ expression is used afterwards is expressed as equation (17), (18), (19) and (20),

$$\hat{\alpha}_1 = \bar{x} = \sum_{i=1}^{n} x_i/n, \tag{17}$$

$$\hat{\alpha}_2 = \hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n, \tag{18}$$

$$\hat{\alpha}_3 = \sum_{i=1}^{n} (x_i - \bar{x})^3/(n\,\hat{\sigma}^3), \tag{19}$$

$$\hat{\alpha}_4 = \sum_{i=1}^{n} (x_i - \bar{x})^4/(n\,\hat{\sigma}^4). \tag{20}$$

The correlation between parameters λ1, λ2, λ3 and λ4 and momentsα1, α2, α3, and α4 is defined using four nonlinear expressions, whose solution is found by numerical equations. But, owing to the complexity of the computational process, the parameters λ1, λ2, λ3, and λ4are considered as functions of α3 and α4 for standardized data where α1=0 and α2 =1.

The basic source distributions are estimated using the marginal distributions by adapting them to the EGLD family employing the method of moments as defined. The inverse distribution function is used to determine the scoring function since the density function D of the EGL is not in a closed form from equation (21).

$$p = F(y) \tag{21}$$

where $F(y)$ refers to the EGLD's distribution function, and the following formula is obtained for the score function expressed as equation (22),

$$\varphi(p) = -\frac{\lambda_2 p^{\lambda_3-2}(\lambda_3-1)\lambda_3}{\left(p^{\lambda_3-1}\lambda_3+(1-p)^{\lambda_4-1}\lambda_4\right)^2} + \frac{\lambda_2(1-p)^{\lambda_4-2}(\lambda_4-1)\lambda_4}{\left(p^{\lambda_3-1}\lambda_3+(1-p)^{\lambda_4-1}\lambda_4\right)^2} \quad (22)$$

The original approach where maximum likelihood contrasts are utilized to optimize the criterion generated expressed as equation (23),

$$W_{k+1} = W_k + \eta(I - \varphi(y)y^T)W_k \quad (23)$$

where equation (24), η stands for the fixed-point algorithm and learning rates.,

$$W_{k+1} = W_k + D(E\{\varphi(y)y^T\} - diag(E\{\varphi(y_i)y_i\}))W_k \quad (24)$$

where D = diag($\frac{1}{E\{\varphi(y_i)y_i\}-E\{\varphi'(y_i)\}}$).

The EGLD-ICA procedure might be as follows: This procedure is repeated until the convergence conditions are met.

1. For the current data, $y_k = W_k x$ compute the third and fourth sample moments, α3 and α4, and choose EGLD if $\alpha_4 > 2.2 + 2 * \alpha_3^2$.

2. estimate the parameters for EGLD using the technique of moments and compute scores $\varphi(y_k)$.

3. Calculate W_{k+1}, the demixing matrix.

Algorithm 1. shows the EGLD-ICA that is being addressed.

Algorithm 1. EGLD-ICA

Input: Number of Attributes

Output: Optimal features
Step 1: Initialization
Step 2: Initialize two scalar-valued random variables $\boldsymbol{y_1 and y_2}$.
Step 3: Compute the found combined numeric data X is, and the mean is subtracted from the observed data set to set it zero mean as $\boldsymbol{X_c \leftarrow X - E\{X\}}$
Step 4: Compute centered data Xc′s covariance matrix cov X.
Step 5: Decompose cov X using eigenvalues
$\boldsymbol{Z = D^{-1/2}E * X_c}$.
Step 6: Calculate distributions // Enhance generalizations of lambda distributions
Compute the third and fourth sample moments $\alpha_3$and α_4 for current data. $\boldsymbol{y_k = W_k x}$and choose EGLD if $\boldsymbol{\alpha_4 > 2.2 + 2 * \alpha_3^2}$.
Estimate the parameters for EGLD using the method of moments and compute scores$\boldsymbol{\varphi(y_k)}$.
Compute the demixing matrix $\boldsymbol{W_{k+1}}$.
Step7 : Select an initial random vector w of unit norm applying distribution function
Step 8: Iterate until the termination criterion
Step 9: End

3.5. Classification using SVM-ELFFOA

SVM is a learning method and supervised learning model used for regression and classification data analysis. Given a set of training examples, each labeled as one of two classes, the SVM training approach creates a model that places a new instance in one of the two classes. This makes the model a non-probabilistic binary linear classifier. The data set gets subdivided into training and test sets. The test set size has to be around 30 to

40% of the total data size. The test set doesn't take part in controlling SVM-classifier parameters. It is just utilized to verify the accuracy achieved with the classifier.

Representation of isolating hyperplanes for objects coming from the training set can be given using expression. $< w, z > +b = 0$, where w refers to a vector in a perpendicular direction to isolating hyperplane, b indicates a parameter that is associated with the shortest distance from coordinates origin to hyperplane $< w, z >$ stands for vectors w and z's scalar product [37]. The criteria $-1 < w, z > +b < 1$ indicates a strip differentiating the classes.

If classes are linearly separable, a hyperplane is selected to ensure that no object from the training set exists between them. Then, the quadratic optimization issue is solved by maximizing the distance between hyperplanes (strip width) $2/< w, w >$ as in expression as equation (25).

$$\begin{cases} < w, w > \rightarrow min, \\ y_i.(< w, z_i > +b) \geq 1, \quad i = 1, \bar{s} \end{cases} \tag{25}$$

The problem involved with the differentiating hyperplane construction can be re-defined as the dual problem of looking for a Lagrange function's saddle point, therefore, with only two variables, is then simplified to a quadratic programming problem:

$$\begin{aligned} &-L(\lambda) = -\sum_{i=1}^{s} \lambda_i + \frac{1}{2}.\sum_{i=1}^{s}\sum_{\tau=1}^{s} \lambda_i . \lambda_\tau . y_i . y_\tau . \kappa(z_i, z_\tau) \\ &\rightarrow \underset{\lambda}{^{min,}} \sum_{i=1}^{s} \lambda_i . y_i = 0, \;\; 0 \leq \lambda_i \leq C, i = 1, \bar{s} \end{aligned} \tag{26}$$

Where equation (26), the dual variable is represented as λ_i, training set object is indicated as z_i, number either -1 or +1 is represented using y_i, from the experimental dataset, object z_i's class is represented using this number, and the kernel function is given by $\kappa(z_i, z_\tau)$, regularization parameter is indicated as C, and it is >0; in the experimental data set, the object quantity is referred to as S, i= 1, S. The kernel function type $\kappa(z_i, z_\tau)$ needs to be decided during the SVM classifier training. Overall error reduction and balance between enhancements in class isolation gap are obtained using the regularization parameter C and kernel parameter values. So, one of the below-mentioned functions is used as a kernel function$\kappa(z_i, z_\tau)$.

These kernel functions allow the division of objects from various groups. After the training of the SVM classifier, support vectors have to be decided. These vectors are nearest to the hyperplane, which differentiates classes and has all the information on isolation between classes. The primary issue while training the SVM classifier is the recommendations deficit to select regularization parameters, type of kernel function, and kernel function parameter values, which, in turn, can produce superior accuracy concerning object classification. This issue can be resolved using the Enhanced Levy Flight-based Fruitfly Optimization Algorithm.

- **Fruitfly Optimization Algorithm (FOA)**

FOA is a meta-heuristic algorithm drawing its inspiration from the foraging aspect of fruitfly. Fruitflies typically depend on their vision and smell capabilities to identify the location of food. FOA imitates the flight of fruitflies to resolve various optimization problems. In the case of FOA, the fruitfly population (candidate solution) is first established randomly. Then, each fruitfly will update its position following its preferred flying style [30, 31, 32]. Through iteration, the fruitfly population continuously improves the population's fitness (quality of solution).

- **Enhanced Levy Flight-based Fruit Fly Optimization Algorithm (ELFFOA)**

Levy's flights (LF) are arbitrary directions in brief steps. This feature is handy, prevention from slipping into local optimum over the whole population [38]. This enhances the algorithm's capability for worldwide detection. For FOA to traverse the search space efficiently, this research incorporates the LF mechanism. The amended position follows the guidelines provided below equation (27).

$$X_i^{levy} = X_i + X_i \oplus levy(s) \tag{27}$$

Where, when an update is completed, X_i^{levy} denotes the updated position of the ith search agent X_i.

This research contribution introduces an enhanced SVM, which uses the ELFFOA mechanism. The produced SVM-ELFFOA model can also provide adaptive decision-making for the two main SVM hyperparameters. There are two essential components to the framework. The outside classification performance analysis is the second, while the inner parameter optimization is the first. Using the 5-fold Cross-Validation (CV) analysis, the ELFFOA technique adaptively adjusts SVM parameters in the inner parameter optimization stage [30, 33, 34, 35]. Then, the SVM prediction model is fed the optimal parameters acquired from the 10-fold CV analysis to carry out the classification task for DM diagnosis in the external loop. The fitness function that was used was classification accuracy.

$$fitness = (\sum_{i=1}^{K} ACC_i)/k \tag{28}$$

where equation (28), ACCi indicates the mean accuracy that the SVM classifier attains.

Algorithm 2. SVM – ELFFOA

Input: Features Count
Output: Classified diabetes data
Step 1: Initialization
Step 2: Get the optimal values to tune SVM model parameters
Step 3: Training of the SVM model
Step 4: $\boldsymbol{p} \leftarrow \boldsymbol{p}^*$
Step 5: while $\boldsymbol{p} \geq \boldsymbol{2}$ do
Step 6: $\boldsymbol{SVM_p} \leftarrow \boldsymbol{SVM}$ Given data observations and p - variable tuning parameters that are optimized
Step 7: Optimization of kernel function // ELFFOA
Initialize the fruitfly population for every fruitfly
Compute the fitness of fruitflies per the reciprocal of the particle's distance from the starting point;
end for
Consider the global optimum as the position of the fruitfly swarm, which is the ideal fruitfly position;
t=0;
while ($\boldsymbol{t \leq Maxnumberofiterations}$) for each fruitfly
Update the current fruitfly's position depending on the fruitfly swarm's position by applying eq. (5.27);
Compute the fitness of fruitflies per the reciprocal distance from the particle to the origin by applying eq. (5.28);
end for
If the best individual in the fruitfly population has a higher fitness than the global optimum, update the global optimum;
t=t+1;
end while

Step 8: $\boldsymbol{w_p} \leftarrow$compute the weight vector of the $\boldsymbol{SVM_p(w_{p1}, \ldots\ldots \boldsymbol{w_{pp}})}$
Step 9: $\boldsymbol{rank.criteria} \leftarrow (\boldsymbol{w_{p1}}^2, \ldots\ldots \boldsymbol{w_{pp}}^2)$
Step 10: min.rank. Criteria$\leftarrow$variable with the least value in rank. Criteria vector
Step 11: discard min.rank. Criteria from data
Step 12:$\boldsymbol{rank_p \leftarrow min.rank.criteria}$
Step 13: $\boldsymbol{p \leftarrow p - 1}$
Step 14: end
Step 15:$\boldsymbol{rank_1 \leftarrow variableindata \notin (rank_2, \ldots\ldots rank_{p*})}$
Step 16: Return$(\boldsymbol{rank_1, \ldots\ldots rank_{p*}})$
Step 17: End

Algorithm 2, shows the Support Vector Machine-based Enhanced Levy Flight-based Fruitfly Optimization Algorithm [39]. Consistent Fuzzy Preference Relations (CFPR) were used for this work because they provide a simple and effective way to deal with decision-making situations where preferences among options must be appropriately expressed. Decision-making reliability and validity are maintained by CFPR-enforced consistent pairwise evaluation of alternatives. Intuitionistic Fuzzy and Neutrosophic methods incorporate additional membership functions to provide more complex representations of uncertainty and hesitation. However, these methods can be computationally burdensome and add unnecessary complexity when the main goal is maintaining consistency in preference relations. When clarity and consistency in decision-making are of utmost importance, CFPR is a great alternative to the more complex representations of uncertainty provided by Intuitionistic Fuzzy or Neutrosophic approaches due to its balance between simplicity and efficacy.

4. RESULTS AND DISCUSSION

Datasets related to diabetic patients are considered in this work, and different methods are applied for performance assessment with indicators such as accuracy, recall, and precision. Based on the results, classifiers are selected. The experiment results are studied using a MATLAB toolkit through a visualized interface, as shown in Table 2. There are four different results in the prediction process, namely false negative (FN), false positive (FP), true negative (TN) and true positive (TP).

The precision is computed as in (29)

$$Precision(Pr) = \frac{TP}{TP+FP} \tag{29}$$

The recall is also termed as sensitivity, and it is expressed as in (30)

$$Recall(Re) = \frac{TP}{TP+FN} \tag{30}$$

Specificity is also termed a true negative rate (TNR). The ratio between correct negative prediction count and total negative count defines specificity (SP). It is expressed as in (31),

$$\text{Specificity } = \frac{TN}{TN+FP} \tag{31}$$

The test's accuracy is measured using the F-measure. For computing the score, recall 're' and precision 'pr' are considered by this. The ratio between correct positive results count, and "Pr" is defined by all positive results that the classifier returns. The ratio

between the correct positive results count, and all relevant samples count defines 're'. F-measure can be defined as in (32),

$$\mathrm{F-measure} = 2.\frac{Pr.Re}{Pr+Re} \tag{32}$$

Precision and sensitivity's geometric mean defines G-mean, and it is expressed as in (33),

$$\mathrm{G-mean} = \sqrt{Sen * Pre} \tag{33}$$

A highly intuitive performance measure is accuracy, and a ratio between properly identified observation and total observations defines this, and it is expressed as in (34),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{34}$$

Table 2: Performance Comparison Metrics vs. Classifiers

Classifiers	Results (%)						
	Sensitivity	Specificity	Precision	F-measure	G-mean	Accuracy	Error
IKM+LR	85.19	84.78	86.79	85.98	84.98	85.00	15.00
IKM-ADWFA+LR	93.75	88.97	89.67	86.33	91.32	90.50	9.50
IKM-EIWBBA+SVM	94.15	89.16	91.45	88.57	92.87	91.87	8.13
IKM-EIWBBA+ SVM-ELFFOA	95.74	91.21	94.84	90.50	95.84	93.50	6.50

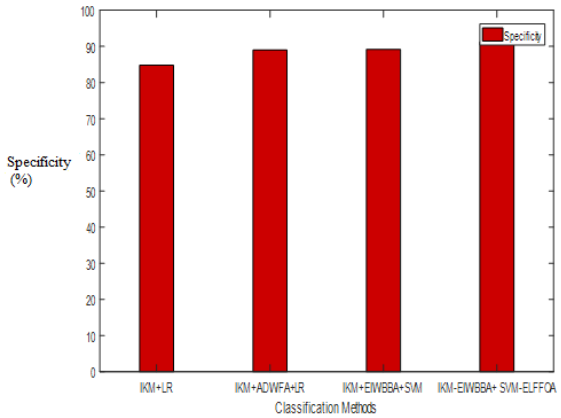

Figure 2: Results of performance comparison between techniques and specificity

Figure 2 illustrates the results of the performance comparison analysis between available IKM+LR, IKM-EIWBBA+SVM, IKM-ADWFA+LR and the discussed IKM-EIWBBA+ SVM-ELFFOA classifier. It can be concluded that the discussed IKM-EIWBBA+SVM-ELFFOA classifier produces a better specificity of 91.21%, while the available IKM-EIWBBA+SVM yields 89.16%, IKM-ADWFA+LR renders 88.97% and IKM+LR gives just 84.78%.

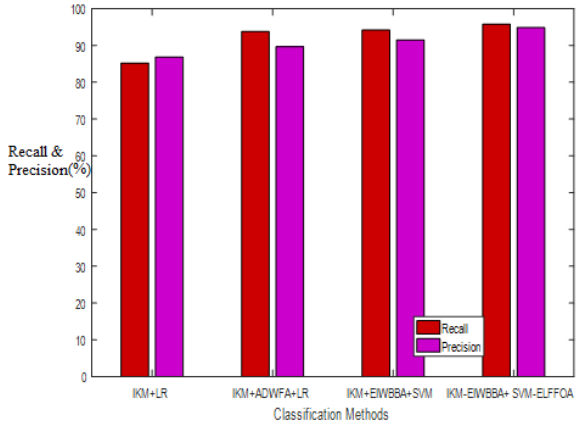

Figure 3: Precision and recall performance comparison results compared to techniques

The precision and recall comparison results of the available IKM+LR, IKM-EIWBBA+SVM, and IKM-ADWFA+LR are shown in Figure 3, along with a discussion of the IKM-EIWBBA+SVM-ELFFOA classifier. The discussed IKM-EIWBBA+SVM-ELFFOA classifier yields much better precision results of 94.84%, while the available IKM-EIWBBA+SVM produces91.45%, IKM-ADWFA+LR renders 89.67% and IKM+LR gives just 86.79%. If the recall is considered, the discussed IKM-EIWBBA+SVM-ELFFOA classifier yields much better recall results of 95.74%, while the available IKM-EIWBBA+SVM produces 94.15%, IKM-ADWFA+LR renders 93.75%, and IKM+LR renders just 85.19%.

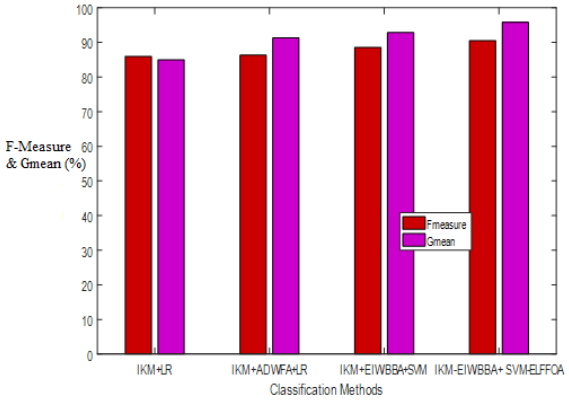

Figure 4: F-Measure and G-Mean performance comparison results compared to approaches

The F-measure and G-mean metric comparison findings between the mentioned IKM-EIWBBA+SVM-ELFFOA classifier and the available IKM+LR, IKM-EIWBBA+SVM, and IKM-ADWFA+LR are shown in Figure 4. It is evident that discussed IKM-EIWBBA+SVM-ELFFOA classifier yields much better G-mean results of 95.84%, while the available algorithm IKM-EIWBBA+SVM produces 92.87%, IKM-ADWFA+LR

gives 91.32% and IKM+LR renders just 84.98%. If F-measure is considered, the discussed IKM-EIWBBA+SVM-ELFFOA classifier yields much better F-measure results of 90.50%, while the available IKM-EIWBBA+SVM yields 88.57%, IKM-ADWFA+LR renders 86.33% and IKM+LR yields just 85.98%.

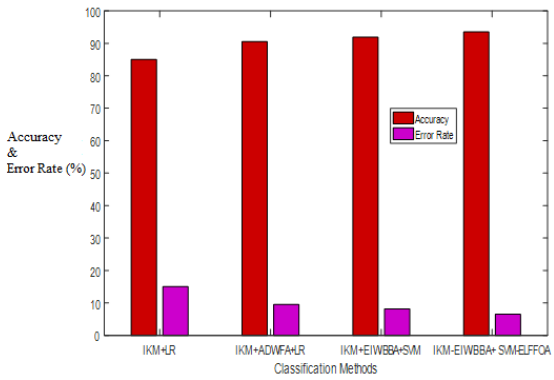

Figure 5: Accuracy vs. techniques performance comparison results

The result accuracy and error rate comparison study between the IKM-EIWBBA+SVM, IKM-ADWFA+LR, and available IKM+LR is shown in Figure 5, along with a discussion of the IKM-EIWBBA+SVM-ELFFOA classifier. It is evident that the discussed IKM-EIWBBA+SVM-ELFFOA classifier produces much better values of the accuracy of 93.50%, while the available IKM-EIWBBA+SVM yields 91.87%, IKM-ADWFA+LR renders 90.50%, and IKM+LR renders just 85.00%. If the error rate is considered, the discussed IKM-EIWBBA+SVM-ELFFOA classifier produces a much-reduced error rate of 6.50% while the available IKM-EIWBBA+SVM yields 8.13%, IKM-ADWFA+LR renders 9.50% and IKM+LR yields just 15.00%.

Patient privacy and data security are paramount when dealing with sensitive medical information. This may be achieved by closely following legal and regulatory requirements, such as HIPAA or the General Data Protection Regulation (GDPR). To avoid breaches and unwanted access, it is crucial to encrypt and anonymize data at every step of processing. Patients' privacy may be better protected by using strong de-identification methods, explicit communication about data usage, and informed permission from patients. Audits and monitoring should be carried out regularly to keep these standards up and deal with any ethical issues as soon as they arise.

5. CONCLUSION AND FUTURE DIRECTION

Early-stage detection plays a major role in diabetes treatment. An ML technique is described in this research work for predicting diabetes. In datasets, this technical work presents an ANFIS for obtaining missing values. Then, an EIWBBA is used to select seeds effectively in the Improved K-means algorithm. In the next stage, the EGLD-ICA-based feature selection technique with minimized time is proposed for dimensionality reduction. At last, the classification is carried out using SVM-ELFFOA. It is evident that the discussed IKM-EIWBBA+SVM-ELFFOA classifier produces much better values of

the accuracy of 93.50%, while the available IKM-EIWBBA+SVM yields 91.87%, IKM-ADWFA+LR renders 90.50%, and IKM+LR renders just 85.00%. From the simulation experiment, the proposed classification techniques implemented in MATLAB software and comparison results indicate that this proposed model has a higher prediction accuracy of 93.50% compared to existing classification methods. In future, information from various locales worldwide can be collected to extend this work for diabetes conclusion and may provide a highly precise and general prescient model. Observing the model's behaviour in a mixed-gender group would be fascinating. Modeling-wise, ELFFOA's hyperparameter optimization may be modified to examine the result and correctness. This technique may be improved and expanded for automated diabetes analysis. Additionally, it can be a useful tool for diabetes educators and practitioners to employ when making treatment decisions that will enhance patients' quality of life. However, this study has a limitation in the variety of training datasets; the model's performance may change when executed on other populations. Factors including genetic variety, lifestyle variances, and regional healthcare practices may impact the accuracy of predictions. These factors might introduce biases or errors when the model is applied to new or unidentified information. Experimenting with the proposed strategy on numerous datasets representing different demographics and data quality levels is crucial for ensuring its robustness and generalizability. Future studies should aim to find ways to test the model in different environments and with diverse healthcare systems.

Conflict of interest statement: No conflicts of interest have been revealed by the author.

Funding: This research received no external funding

REFERENCES

[1] D. Atlas, International diabetes federation. *IDF Diabetes Atlas, 7th ed. Brussels, Belgium: International Diabetes Federation*, vol. 33, no. 2, 2015.

[2] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, 37(Supplement_1), pp. S81-S90, 2014. doi: https://doi.org/10.2337/dc14-S081

[3] R. Kahn, "Follow-up report on the diagnosis of diabetes mellitus: the expert committee on the diagnosis and classifications of diabetes mellitus," *Diabetes care*, vol. 26, no. 11, pp. 3160, 2003.

[4] W. Kerner and J. Brückel, "Definition, classification and diagnosis of diabetes mellitus," *Experimental and clinical endocrinology & diabetes*, vol. 122, no. 07, pp. 384-386, 2014. doi: https://doi.org/ 10.1055/s-0034-1366278

[5] J. P. Kandhasamy and S. J. P. C. S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus." *Procedia Computer Science*, vol. 47, pp. 45-51, 2015. doi: https://doi.org/10.1016/j.procs.2015.03.182

[6] 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, pp. 515, 2018. doi: https://doi.org/10.3389/fgene.2018.00515

[7] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 37, no. 1, pp. S81-S90, 2014. doi: https://doi.org/10.2337/dc14-S081

[8] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962-969, 2017. doi: https://doi.org/10.1016/j.ophtha.2017.02.008

[9] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan and J. Dong, "Identifying medical diagnoses and

treatable diseases by image-based deep learning." *Cell*, vol. 172, no. 5, pp. 1122-1131, 2018. doi: https://doi.org/10.1016/j.cell.2018.02.010

[10] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578-1585, 2018. doi: https://doi.org/10.1016/j.procs.2018.05.122

[11] 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, no. 9, pp. 871-878, 2018.

[12] M. Mounika, S. D. Suganya, B. Vijayashanthi and S. K. Anand, "Predictive analysis of diabetic treatment using classification algorithm," *Int J Comput Sci Inf Technol*, vol. 6, pp. 2502-2502, 2015.

[13] A. Pavate and N. Ansari, "Risk prediction of disease complications in type 2 diabetes patients using soft computing techniques," In *2015 Fifth International Conference on Advances in Computing and Communications (ICACC),* pp. 371-375, 2015. doi: https://doi.org/10.1109/ICACC.2015.61

[14] A. Kumar Dewangan and P. Agrawal, "Classification of diabetes mellitus using machine learning techniques," *International Journal of Engineering and Applied Sciences*, vol. 2, no. 5, 2015.

[15] K. Saravananathan and T. Velmurugan, "Analyzing diabetic data using classification algorithms in data mining," *Indian Journal of Science and Technology*, vol. 9, no. 43, pp. 1-6, 2016. doi: https://doi.org/ 10.17485/ijst/2016/v9i43/93874

[16] R. Joshi and M. Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," *International Research Journal of Engineering and Technology*, vol. 4, no. 10, 2017.

[17] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big data*, vol. 6, no. 1, pp. 1-19, 2019. doi: https://doi.org/10.1186/s40537-019-0175-6

[18] S. Hina, A. Shaikh and S. A. Sattar, "Analyzing diabetes datasets using data mining," *Journal of Basic & Applied Sciences*, vol. 13, pp. 466-471, 2017. doi: https://doi.org/10.6000/1927-5129.2017.13.77

[19] R. Asgarnezhad, M. Shekofteh and F. Z. Boroujeni, "Improving Diagnosis of Diabetes Mellitus Using Combination of Preprocessing Techniques," *Journal of Theoretical & Applied Information Technology*, vol. 95, no. 13, 2017.

[20] C. C. Jorge, H. N. Jorge, M. R. Wendell, S. A. Edalatpanah, A. B. Shariq, S. Naz, J. C. Javier and P. E. Gabriel, "Novel characterization and tuning methods for integrating processes," *International Journal of Information Technology*, vol. 16, no. 3, pp. 1387-1395, 2024. doi: https://doi.org/10.1007/s41870-023-01679-9

[21] M. S. Farahani, H. Farrokhi-Asl and S. Rahimian, "Hybrid Metaheuristic Artificial Neural Networks for Stock Price Prediction Considering Efficient Market Hypothesis," *International Journal of Research in Industrial Engineering,* vol. 12, no. 3, pp. 2783-1337, 2023. doi: htttps://doi.org/ 10.22105/riej.2023.361216.1336

[22] M. Lincy Jacquline and N. Sudha, "Weighted fuzzy C means and enhanced adaptive neuro-fuzzy inference based chronic kidney disease classification," *Journal of Fuzzy Extension and Applications*, vol. 5, no. 1, pp. 100-115, 2024. doi: https://doi.org/10.22105/jfea.2024.437690.1376

[23] R. Rasinojehdehi and S. E. Najafi, "Advancing risk assessment in renewable power plant construction: an integrated DEA-SVM approach," *Big Data and Computing Visions*, vol. 4, no. 1, pp. 1-11, 2024. doi: https://doi.org/10.22105/bdcv.2024.447876.1178

[24] M. Mohamadjkhani, R. Radfar, N. Pilevarisalmasi and M. Afsharkazemi, "Selection of open innovation method in the automotive industry using Adaptive network-based fuzzy inference system (ANFIS)," *Journal of Applied Research on Industrial Engineering,* vol. 11, no. 4, 2024. https://doi.org/10.22105/jarie.2023.393194.1543

[25] S. Alam, J. Kundu, S. Ghosh and A. Dey, "Trusted fuzzy routing scheme in flying ad-hoc

network," *Journal of Fuzzy Extension and Applications*, vol. 5, no. 1, pp. 48-59, 2024. doi: https://doi.org/10.22105/jfea.2024.436052.1370

[26] A. A. El-Douh, S. Lu, A. Abdelhafeez, A. M. Ali and A. S. Aziz, "Heart Disease Prediction under Machine Learning and Association Rules under Neutrosophic Environment," *Neutrosophic Systems with Applications*, vol. 10, pp. 35-52, 2023. doi: https://doi.org/10.61356/j.nswa.2023.75

[27] N. Khalil, M. Elkholy and M. Eassa, "A Comparative Analysis of Machine Learning Models for Prediction of Chronic Kidney Disease." *Sustainable Machine Intelligence Journal*, vol. 5, pp. 3-1, 2023. doi: https://doi.org/10.61185/SMIJ.2023.55103

[28] A. M. Ali and S. Broumi, "Machine Learning with Multi-Criteria Decision-Making Model for Thyroid Disease Prediction and Analysis," *Multicriteria Algorithms with Applications*, vol. 2, pp. 80-88, 2024. doi: https://doi.org/10.61356/j.mawa.2024.26961

[29] S. Mandour, A. Gamal and A. Sleem, "Mantis Search Algorithm Integrated with Opposition-Based Learning and Simulated Annealing for Feature Selection," *Sustainable Machine Intelligence Journal*, vol. 8, pp. 5-56, 2024. doi: https://doi.org/10.61356/SMIJ.2024.8300

[30] H. Wu, S. Yang, Z. Huang, J. He and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100-107, 2018. doi: https://doi.org/10.1016/j.imu.2017.12.006

[31] G. Krishnaveni and T. Sudha, "A novel technique to predict diabetic disease using data mining–classification techniques," *International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS)*, vol. 3, 2017.

[32] S. Alby and BL Shivakumar, "A prediction model for type 2 diabetes using adaptive neuro-fuzzy interface system," Biomedical Research, Special Issue: Computational Life Sciences and Smarter Technological Advancement: Edition: II, pp. 69-74, 2018. doi: 10.4066/biomedicalresearch.29-17-254

[33] S. Alby and B. L. Shivakumar, "A prediction model for type 2 diabetes risk among Indian women," *ARPN Journal of Engineering and Applied Sciences*, vol. 11, no. 3, pp. 2037-2043, 2016.

[34] X. Huang, X. Zeng and R. Han, "Dynamic inertia weight binary bat algorithm with neighborhood search," *Computational intelligence and neuroscience*, vol. 2017, no. 1, pp. 3235720, 2017. doi: https://doi.org/10.1155/2017/3235720

[35] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411-430, 2000. doi: https://doi.org/10.1016/S0893-6080(00)00026-5

[36] O. Chapelle, V. Vapnik, O. Bousquet and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine learning*, vol. 46, pp. 131-159, 2002. doi: https://doi.org/10.1023/A:1012450327387

[37] W. T. Pan, "A new fruit fly optimization algorithm: taking the financial distress model as an example," *Knowledge-Based Systems*, vol. 26, pp. 69-74, 2012. doi: https://doi.org/10.1016/j.knosys.2011.07.001

[38] M. A, Kumar and I. L. Aroquiaraj, "Adaptive Divergence Weight Firefly Algorithm (ADWFA) with Improved K-Means Algorithm and Adaptive Neuro Fuzzy Inference System (ANFIS) for Type 2 Diabetes Mellitus Prediction," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, no. 6, pp. 18-31, 2009.

[39] M. P. R. Ganesan, "Hybrid Genetic Discretization model with Parental comparison using Correlation Clustering for Distributed DNA Databases," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 5, 2022.