Utilizing Multimodal Medical Data and a Hybrid Optimization Model to Improve Diabetes Prediction

A. Leela Sravanthi[1*], Sameh Al-Ashmawy[2], Dr. Chamandeep Kaur[3],
Dr. Mohammed Saleh Al Ansari[4], Dr. K. Aanandha Saravanan[5], Dr. Veera Ankalu. Vuyyuru[6]
Assistant Professor, Department of Information Technology, Marri Laxman Reddy Institute of Technology and Management, Dundigal, Hyderabad-500043, India[1]
Imam AbdulRahman Bin Faisal University, Kingdom of Saudi Arabia, and Damanhour University, Egypt[2]
Lecturer, Department of CS & IT, Jazan University, Saudi Arabia[3]
Associate Professor, College of Engineering-Department of Chemical Engineering, University of Bahrain, Bahrain[4]
Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology[5]
Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522502, A.P, India[6]

Abstract*—Diabetes is a major health issue that affects people all over the world. Accurate early diagnosis is essential to enabling adequate therapy and prevention actions. Through the use of electronic health records and recent advancements in data analytics, there is growing interest in merging multimodal medical data to increase the precision of diabetes prediction. In order to improve the accuracy of diabetes prediction, this study presents a novel hybrid optimisation strategy that seamlessly combines machine learning techniques. In order to merge many models in a way that maximises efficiency while enhancing prediction accuracy, the study employs a collaborative learning technique. This study makes use of two separate diabetes database datasets from Pima Indians. A feature selection process is used to streamline error-free classification. A third method known as Binary Grey Wolf-based Crow Search Optimisation (BGW-CSO), which was produced by merging the Binary Grey Wolf Optimisation Algorithm (BGWO) and Crow Search Optimisation (CSO), is provided to further enhance feature selection capabilities. This hybrid optimisation approach successfully solves the high-dimensional feature space challenges and enhances the generalisation capabilities of the system. The Support Vector Machine (SVM) method is used to analyse the selected characteristics. The performance of conventional SVMs is enhanced by the newly created BGW-CSO technique, which optimises the number of hidden neurons within the SVM. The proposed method is implemented using Python software. The suggested BGW-CSO-SVM approach outperforms the current methods, such as Soft Voting Classifier, Random Forest, DMP_MI, and Bootstrap Aggregation, with a remarkable accuracy of 96.62%. Comparing the suggested BGW-CSO-SVM approach to the other methods, accuracy shows an average improvement of around 16%. Comparative evaluations demonstrate the suggested approach's improved performance and demonstrate its potential for real-world use in healthcare settings.

Keywords*—Diabetes prediction; multimodal medical data; binary grey wolf optimization; crow search optimization; support vector machine

I. Introduction

Diabetic is a long-term endocrine illness that alters the structure of the human body and impacts metabolism. From 100 million to 422 million people, the illness has expanded more since 2014 [1]. Excessive blood sugar levels brought on by inadequate insulin production or releases are the main contributing factor to diabetes, a metabolic illness. In 2010, it was estimated that 285 million individuals worldwide will have diabetes. By 2030, this number will rise to 552 million based on the disease's present rate of progression. By 2040, it is anticipated that one in ten persons would develop diabetes [2]. Due to varying behaviors, lifestyles, and living standards, diabetes is becoming increasingly common. Therefore, it is important to do research on how to accurately and quickly diagnose and treat diabetes. Diabetes is an extremely dangerous chronic illness that is preventable. The likelihood of developing diabetes is expected to drastically increase during the next 50 years. Diabetic is a condition that impairs function due to inadequate levels of insulin in the bloodstream. Indications of hyperglycemia might include increased appetite, thirst, and frequency of urine [3]. The three primary kinds of diabetes that exist are Type 1, Type 2, and gestational diabetes. Type 2 diabetes is becoming more common, and it constitutes one of the leading causes of death worldwide. The absence of insulin in the human body affects individuals despite their age or gender [4] [5]. Type 1 diabetes develops as a result of a shortage of insulin. Instead of protecting the human body against harmful viruses or bacteria, the immune system attacks and destroys the cells that make insulin in the pancreas.

The professionals believe that both inherited and ecological circumstances have a substantial impact on the condition, despite the fact that the underlying cause of diabetes is still unknown. Although not curable, it may still be managed with treatment and medication [6] [7]. People with diabetes face the risk of developing additional medical issues such as heart disease and damaged nerves. Thus, by avoiding difficulties, early detection and treatment of diabetes can lower the chance of developing major health problems [8] [9]. The sole treatment for this kind of diabetes is to supply the patient's body by injecting the necessary quantity of insulin [10]. The pancreas has no ability to generate enough insulin to overcome this barrier whenever an individual develops

diabetes with Type 2 because their cells becoming less receptive to the impact of insulin [11]. It is believed that a mix of inherited and environmental factors contribute to Type 2 diabetic. Diabetes with Type 2 is largely associated with being overweight. Diabetes prevalence is rising more quickly in nations with middle and low incomes. Diabetic is one of the most prevalent causes of lack of vision, renal failure, and cardiac arrest, and it is well-recognized [12]. Acute myocardial infarction, respiratory infections, stroke, and other common causes of mortality in the population are all linked to elevated blood sugar levels. Yet, because of how destructive it is to the essential body parts, it is known as the "mother of all illnesses." The majority of women who suffer from diabetes are unaware of their condition, reported to the World Health Organisation (WHO). The condition can spread to kids, particularly among pregnant women. In addition to additional chronic and fatal illnesses, diabetic women are in danger of premature delivery, renal failure, heart attacks, lack of vision, and other conditions. Determining diabetic in pregnant women as quickly as feasible is therefore crucial [13]. In India, 32 million persons had diabetes overall in the year 2000. The Diabetes Epidemic in India is explored in Fig. 1. The number rose to 41 million people in 2007, then to 62 million in 2011, and finally to 73 million in 2017. By the decade 2045, 134 million more individuals are anticipated to be living in this situation.

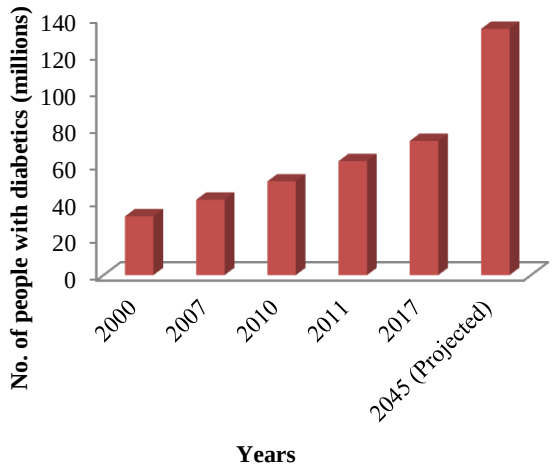

Fig. 1. Diabetes Epidemic in India during the year of 2000 to 2045 [14].

For more precise and timely prediction, an advanced medical framework for suggestions is becoming ever more important daily. In order to lower the likelihood that these diseases will strike individuals, research call for a system that is effective at spotting and effectively treating life-threatening ailments like diabetes. Forecasting accurately an individual's risks for the disease may facilitate personalized medical treatment and wellness management strategies as well as give healthcare system decision-makers an overview of upcoming illness risk variations in the community, allowing them to develop plans for the provision of associated healthcare services and lowering the burden of illness on society as a whole. Machine learning methods are employed in many sectors, and they have a successful association with the medical field. It is utilized to reduce the price of diagnostics as well as to increase diagnostic accuracy. According to the process of learning, it is difficult to identify a condition from a medical text since the information is unstructured [15]. In order to investigate the predictive modelling of various illness risks for patients, this research uses SVM systems.

The following are the study's main contributions.

- The combination of an extensive range of multimodal medical data, including data analytics and electronic health records, makes a substantial contribution to the study. By adding this, improve the accuracy of diabetes prediction and offer a more thorough knowledge of the variables affecting the condition.
- This work presents a novel hybrid optimization method that integrates the Binary Grey Wolf Optimization Algorithm (BGWO) with the Crow Search Optimization (CSO). This revolutionary approach aims to improve diabetes projections' accuracy and reliability, thereby making a significant contribution to the field of medical condition prediction modelling.
- The use of ensemble learning techniques in the research advances diabetes prediction tools. The goal of the research is to optimize diabetes prediction accuracy by utilizing the advantages of several different models. In the context of intricate medical data, this method offers a useful viewpoint for enhancing the accuracy of predictive models.
- The study makes an important contribution by effectively employing the Support Vector Machine (SVM) mechanism to the analysis of medical time-series data. This application improves the prediction abilities of the model, especially when handling complex temporal patterns found in medical data. Predictive modelling and medical data analysis could benefit from the methodological improvement represented by the use of SVM.

Following is the structure for the remaining portion of the paper: In Section II relevant work based on different approaches for diabetes prediction is discussed, and in Section III and Section IV, the procedure of feature selection and classification for the suggested method is described. Section V addresses the results and debates, and Section VI accomplishes by outlining the future's potential use.

II. Related Works

C. Zhu, Idemudia, and Feng [16] demonstrates that Utilising the PID database, the data mining-based technique looks toward early diabetic diagnosis and prediction. Despite the fact that K-means is simple and appropriate for a wide range of data kinds, it is highly dependent on the original positions of clustering centres that characterize the ultimate clustering result, that either yields a sufficient and effective organised data set for the method of logistic regression or offers a lesser quantity of data as a result of incorrect accumulating of the first set of information, limiting the efficacy of the logistics regression. The major objective was to identify methods for enhancing the reliability of the k-means

cluster and logistical regression results. This framework uses the logistic regression technique, k-means, and PCA. In comparison to the outcomes of other previously reported research, the experimental findings demonstrates that PCA improved the performance of the logistic regression and k-means clustering method classification system, with a k-means outcome that included 25 more properly categorized information and a logistic regression reliability of 1.98% greater. As a result, it is demonstrated that the framework may be effective for autonomously forecasting diabetes utilizing information from patient electronic health records. Due to clustering's complexities and incapacity of recovering from database damage, this approach is ineffective.

D. Wang et al. [17] examines how having diabetes raises your chance for renal failure and significant consequences which includes heart disease. If this condition is detected and treated right away, individuals can live better and have a higher quality of life. Many supervised machine-learning techniques that have been created and trained on relevant datasets can assist in the early detection of this disease. The objective of this research is to create efficient machine-learning-based classification approaches for diagnosing diabetes in individuals using medical data. In this study, the following machine-learning techniques will be taught using various datasets. The investigation has employed label-encoding and normalization as two efficient techniques for pre-processing to boost system dependability. Investigations have also identified and ranked other categories of risk factors using a range of selection of features approaches. The model's effectiveness has been examined through a number of experiments using two different datasets. The results show that, depending on the information's sources and the ML technique utilized, the recommended methodology can deliver higher accuracy with values ranging from 2.71% to 13.13% when the recommended structure is compared to other recent research. The research implements this framework into a web application using the Python Flask web development environment. The findings of this work imply that suitable pre-processing pipelines on medical information and the use of ML-based categorization may correctly and effectively predict diabetic. Due to a small quantity of data set, this strategy is ineffective [18].

Cappon et al. [19] presents the research paper titled "Individualized Models for Glucose Prediction in Type 1 Diabetes: comparing black-box approaches to a physiological white-box one" addresses the critical need for accurate blood glucose prediction in Type 1 diabetes management. The study compares black-box models commonly used for glucose prediction. By individualizing the physiological model through a Bayesian approach, the paper explores the efficacy of different prediction techniques, including non-parametric models, deep learning methods (LSTM, GRU, TCN), and a recursive autoregressive model with exogenous input (rARX). Results demonstrate that black-box strategies, particularly non-parametric models, outperform the personalized white-box model across various prediction horizons. Despite the physiological model's individualized parameters, the study highlights the continued preference for black-box approaches in glucose prediction. These findings contribute valuable insights for the development of next-generation tools in T1D management and decision support systems, emphasizing the importance of optimizing predictive accuracy for patient care and treatment planning.

Xie and Wang [20] offers the research paper titled "Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison with Classical Time-Series Models" intends to evaluate the effectiveness of several artificial intelligence models to anticipate blood glucose concentrations in Type 1 diabetes individuals using time-series data versus a traditional Auto regression using an Exogenous input model. The study analyses various input characteristics, regression model ordering, and prediction techniques to assess ML-based regression models, particularly deep learning models like LSTM and TCN. The performance measures for determining the likelihood of false alarms on hypo/hyper glycemia occurrences include RMSE, chronological gain, and normalised efficiency of second-order differentiation. The outcomes show that for both prediction approaches, the ARX model gets the lowest absolute RMSE, while ML models do not exhibit a significant advantage over the classic ARX model, except for TCN's robustness in handling BG trajectories with spurious oscillations. The study offers insightful information that will help researchers and medical professionals choose the best algorithms for BG predictions in T1D. However, it suggests that ML models do not outperform the ARX model, highlighting the importance of considering the context and characteristics of the data when choosing prediction models for diabetes management.

M. Alirezaei et al. [21] explores that data analysts look at diabetes mellitus for a variety of causes, including the serious health issues that might arise for those who have it, the financial burden it places on healthcare systems, and so on. Investigators examine the patient's lifestyle, genetic data, etc. to determine the primary causes of this illness. Finding patterns that facilitate quick identification of the illness and appropriate therapy is the aim of data mining in this situation. The supply of the proposed treatment technique quickly became nearly impossible because of the large number of information associated with therapeutic settings and illness diagnostics. This supports the application of pre-processing methods and information reduction strategies in these situations. Clusters and meta-heuristic techniques continue to play crucial roles in this area. In this study, outliers are initially recognized and eliminated using a technique depending on the k-means clustering technique. The selection of the least significant traits with the greatest categorization efficacy is then made using four bi-objective meta-heuristic procedures. This is accomplished using SVM, a form of ML technique. Utilizing the tenfold cross-validation method, the constructed model is also verified. This method is ineffective because it works inadequately with large data sets and when the data set contains extra noise.

Diabetes can result from either one of two causes, including insufficient insulin synthesis or insufficient cell sensitivity to the effects of insulin. The purpose of this inquiry is to locate diabetes mellitus using data mining approaches. Numerous techniques can be used to diagnosis diabetic. Data

mining techniques are one method. Health information has benefited from the usage of methods for data mining, leading to some important, effective advancement that can help clinicians make the best decisions. This study suggests a combined method for identifying diabetes using the Sequential Minimal Optimisation (SMO) classifier technique and the FF clustering approaches. The FF clustering technique is used to separate the data into a number of groups. Computation time was greatly reduced as a result of the dataset's reduced dimensionality. The clustered result is sent into the SVM-based classifier. It accurately divides individuals into diabetes and non-diabetic groups, or confirmed negatives and positives. 768 diabetes patient specimens from the Pima Indians Dataset are included in the information set utilized in the identification of diabetics. The trial's results showed that a hybrid data mining strategy could help doctors make better clinical decisions concerning diabetic diagnosis for patients. This approach is ineffective since it is inappropriate for usage with huge datasets [22].

The literature reviews address several approaches to diabetes prediction and diagnosis, each with unique difficulties. A method based on data mining that employs PCA, k-means, and logistic regression. The effectiveness of logistic regression may be impacted by k-means' reliance on initial grouping positions, which might result in errors. greater accuracy but limited by small dataset when using supervised machine learning for diabetes detection. Black-box models for Type 1 diabetes glucose prediction, demonstrate the superiority of non-parametric models over customized white-box models. Machine learning techniques for the prediction of blood glucose, showing that ML models does not perform appreciably better than conventional time-series models and highlighting the significance of taking data features into account. When dealing with enormous datasets and noise, data mining and meta-heuristic approaches are ineffective for diagnosing diabetes. Although FF clustering and the SMO classifier work well together, large datasets are not well served by this combination approach while data volume, model dependence, and contextual factors present difficulties, this research offer valuable insights into the diagnosis and prognosis of diabetes overall.

III. Problem Statement

From the aforementioned debate, the literature review tackles the urgent need for more precise and timely diabetes prediction, a worldwide health issue. Although electronic health records and multimodal medical data are readily available, it's possible that current approaches aren't completely using this rich data source for accurate diabetes diagnosis. The main problem is to increase the model's generalisation capabilities and get over the constraints put on us by high-dimensional feature spaces. In order to increase the precision of diabetes prediction, this work intends to create a novel hybrid optimisation model that smoothly combines machine learning methods, notably Support Vector Machine. Support Vector Machines have a reputation for handling high-dimensional data well and for being able to locate the best hyperplanes to divide various classes in large datasets. This is especially helpful for predicting diabetes because medical data frequently contains a wide range of factors. SVM is a good option due to its resilience in handling such data and its ability to clearly distinguish between instances that are diabetic and those that are not [23].

IV. Proposed BGW-CSO-SVM Approach

For precise and reliable diabetes prediction, this study suggests a unique hybrid optimization approach that incorporates Machine learning. Using a hybrid optimization model and multimodal medical data, the analysis techniques for enhancing diabetes prediction form an all-encompassing strategy. Z-score normalization is used to standardize features across various modalities and fill in missing values in the data as part of the first pre-processing processes. After that, the BGW-CSO algorithm—a hybrid optimization model that includes Crow Search Optimization, Binary Grey Wolf Optimizer—is used for feature selection. To improve model efficiency, this stage seeks to determine which attributes are most pertinent. After that, the features that were chosen are used in the classification stage, when the Support Vector Machine (SVM) method is applied. Utilizing the advantages of feature selection, pre-processing, and classification, the overall strategy combines these techniques to produce a strong framework for diabetes prediction, maximizing the hybrid model's performance on multimodal medical data. Fig. 2 shows how the suggested technique is presented.

A. Data Collection

Databases 1 and 2 were the two databases utilized to test the forecast of Type 2 diabetes. Although the more recent information was gathered through the Mendeley Data website, the earlier data was obtained via the Kaggle website. Database 1, which was first compiled by the National Institute of Diabetes and Digestive and Kidney Diseases in 2004, contains all female patients with diabetes who are at least twenty-one years old and of Pima Indian descent. Following the value that was missing is removed; the database has 392 entries and eight variables, comprising their ages, pregnancy, blood pressure, skin elasticity, the sugar glucose, and insulin levels. The second data set was obtained from the labs of Medical City Hospital and the Specializes Centre for Endocrinology and Diabetes-Al-Kindy Teaching Hospital and is originating from the Iraqi society. Age, urea, Hemoglobin A1c (HBA1C), creatinine ratio, high-density lipoprotein (HDL), cholesterol, very-low-density lipoprotein (VLDL), triglycerides, body mass index (BMI) and low-density lipoprotein (LDL) are the 10 parameters included in this set of information that make up the data characteristic utilized in this investigation. Only information from the diabetic and non-diabetic classes was used to make predictions for Type 2 diabetes. For this investigation, 392 records from this dataset—which is identical to Dataset 1—were chosen at randomness [24].

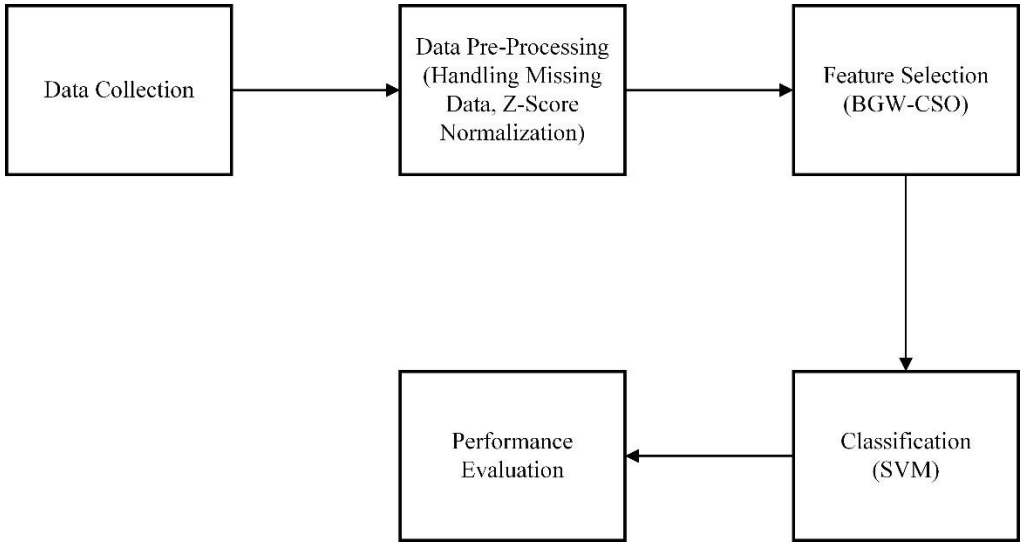

Fig. 2. Proposed methodology.

B. Pre-processing using Handling Missing Data and Z-Score Normalization

The pre-processing phase in the proposed architecture comprises the removal of outliers (S), filling in for missing data (M), and standardisation (R), which are simply defined as follows: The outlier is a data point that differs noticeably from other occurrences. Considering the classifiers are particularly dependent on the data range and the distribution of the characteristics, they need to be excluded from the distribution of the data. This literature's outlier rejection quantitative formulation may be expressed as in Eq. (1).

$$S(y) = \begin{cases} y, \ if M_1 - 1.5 \times IMR \leq x \leq M_3 + 1.5 \times IMR \\ reject, \ Otherwise \end{cases} \quad (1)$$

In this case, y represents the specific occurrences of the feature vectors that lie in n-dimensional space, where y $\in R_n$. The first, third, and interquartile ranges of the characteristics are designated as M1, M3, and IMR, correspondingly, where M1, M3, and IMR $\in R_n$.

After the outliers were eliminated, the attributes were processed to fill in any missing as well as null values because they may cause any classifier to make an incorrect prediction. Instead of dropping, the suggested framework substituted the deficient or null values with the average values of the characteristics, which may be expressed as in Eq. (2). The use of the mean for imputation is advantageous since it attributes continuous information without adding outliers.

$$M(y) = \begin{cases} mean(y), if \ y = \frac{Null\ value}{Missed\ value} \\ y, otherwise \end{cases} \quad (2)$$

Where y represents for the feature vector occurrences in n-dimensional space, which are represented by the expression y belongs to Rn. The process of rescaling the characteristics to create a conventional distribution that is normal with a zero mean and a unit variance is known as standardisation, sometimes known as Z-score normalisation. The data distribution's skewness is likewise lessened by standardisation (R), as seen in Eq. (3).

$$R(y) = \frac{y - \bar{y}}{\sigma} \quad (3)$$

where, y is the n-dimensional instances of the feature vector, y $\in R_n$. $\bar{y} \in R_n$ and $\sigma \in R_n$ are the mean and standard deviation of the attributes.

C. Feature Selection using BGW-CSO

Utilize the BGW-CSO algorithm for feature selection processes. In order to improve the efficacy and efficiency of the ensuing classification model, this entails picking the most pertinent features from the multimodal medical data. In diabetes forecasting, the selection of features is the method of selecting a subset of important traits from the initial information to build an effective predicting model. One can find the most useful and discriminating traits that have a big impact on predicting whether diabetes will exist or not. In diabetic forecasting, selecting features aims at enhancing modelling precision, reducing the dimension, enhancing interpretability, streamlining calculation, and promoting generalization to novel information. Selecting the most relevant factors helps to construct accurate and efficient models for prediction for the assessment, evaluation of risk, and scheduling of diabetes medication. BGW-CSO chooses the features in this case.

1) Binary Grey Wolf Optimization: GWO seems to be a reliable optimization technique. It imitates the harmonious, well-defined interactions at work present in grey wolf eating behavior. Grey wolves frequently live in packs of five to twelve individuals that are rigidly structured under the strong command of the wolf. The predation method used by the GW squad consists of three steps: hunting, encircling, and killing. The leader of the pack was usually the most notable wolf, also known as α wolf. β Wolf and δ wolf, respectively, are the GWO terms for the second and third tiers of leading wolves. These auxiliary wolves, which are second and third in rank, assist the lead wolf in making hunting decisions. The other wolves there are all recognized as ω wolves, and they use these powerful wolves to chase and kill the prey.

In hunts, the encircling of victim's tactic is used. The following Eq. (4) and Eq. (5) for repetition *x* shows how this technique makes sense,

$$\vec{F} = \left|\vec{Z} \times \overrightarrow{A_b}(x) - \vec{P}(x)\right| \quad (4)$$

$$\vec{A}(x+1) = \overrightarrow{A_b}(x) - \vec{F}.\vec{H} \quad (5)$$

In this case, $\vec{F}$ and $\vec{Z}$ are effective factors, which is denoted by the formula $\vec{F} = 2\vec{e}.\overrightarrow{v_1} - \vec{e}$ and $\overrightarrow{Z} = 2.\overrightarrow{v_2}$. Where the randomized vectors $\overrightarrow{v_1}, \overrightarrow{v_2}\ \varepsilon\ (0,1)$ and $\overrightarrow{e} = e_1(1 - {}^{x}/_{maxx})$, gradually decline from e_1 to zero; the value of e_1 was set as 2 in the real GWO. *maxx* also means as many repeats as is feasible. The GWO's top three hunting alternatives have been $\propto$ wolves, β wolves, and γ wolves. The three best possibilities' positions have therefore been kept in the group, and the remaining ω wolves have changed their locations to reflect them. Eq. (6) illustrates the simulation theorem for this location update approach.

$$\vec{A}(x+1) = {\left(\overrightarrow{A_1} + \overrightarrow{A_2} + \overrightarrow{A_3}\right)}/{3} \quad (6)$$

Where, $\overrightarrow{A_1}, \overrightarrow{A_2}, and\ \overrightarrow{A_3}$ is evaluated by Eq. (7),

$$\begin{aligned} \overrightarrow{A_1} &= \overrightarrow{A_\propto}(x) - \overrightarrow{F_1}.\overrightarrow{H_\propto} \\ \overrightarrow{A_2} &= \overrightarrow{A_\beta}(x) - \overrightarrow{F_1}.\overrightarrow{H_\beta} \\ \overrightarrow{A_3} &= \overrightarrow{A_\gamma}(x) - \overrightarrow{F_1}.\overrightarrow{H_\gamma} \end{aligned} \quad (7)$$

Here, $\overrightarrow{H_\propto}$, $\overrightarrow{H_\beta}$, and $\overrightarrow{H_\gamma}$ are evaluated by Eq. (8),

$$\begin{aligned} \overrightarrow{H_\propto} &= \left|\overrightarrow{Z_1} \times \overrightarrow{A_\propto}(x) - \vec{A}\right| \\ \overrightarrow{H_\beta} &= \left|\overrightarrow{Z_2} \times \overrightarrow{A_\beta}(x) - \vec{A}\right| \\ \overrightarrow{H_\gamma} &= \left|\overrightarrow{Z_3} \times \overrightarrow{A_\gamma}x) - \vec{A}\right| \end{aligned} \quad (8)$$

In simple terms, feature selection is a binary issue since the bounds of the feature selection space for searching are zero and one. In the basic GWO method, the process of searching space is ongoing. Consequently, using native GWO to solve the feature selection challenge is not an option. It is necessary to create a modified (binary) variant of the method. To modify the positions of the agents searching for the method known as GWO in the binary searching space, the subsequent sigmoid function was added in Eq. (9):

$$A_{binary}(t+1) = g(a) = \begin{cases} 1, & sigmoid\ \left(\frac{A_1+A_2+A_3}{3}\right) \\ & \geq x \\ 0, & otherwise \end{cases} \quad (9)$$

From the uniform distribution ∈ [0, 1], a random variable *x* is obtained. Here, the binary updating mechanism is $A_{binary}(t+1)$. The sigmoid function *S(a)* is given in Eq. (10).

$$\text{S(a)} = \frac{1}{1 + e^{(-10*(a-0.5))}} \quad (10)$$

2) Crow search algorithm: Askazadeh proposed the Crow search algorithm (CSA), a method with natural form cues [25]. The method is used in based on population evolutionary computation models crow birds' behavior and social relationships. Undoubtedly intelligent creatures with larger-than-average neurons, crows are clever. They dwell in groups called flocks and hide their food in places that may still be found and retrieved a few months later. Additionally, they experience self-consciousness when carrying out the mirror test. They are able to recall appears, and if a poor one is noticed, they can converse with one another incomprehensibly to alert the other crows. Similar to other social animals, crows occasionally steal by carefully determining where other crows store their meals and then taking them. Whenever a crow suspects that someone is tracking it in an effort to deceive a thief, it is going to a new position that is distant from where the food is.

There are N solutions in the population, and the task's parameters are (overall of crow). The vector $l_j^t = \left[l_{j1}^t, l_{j2}^t, \ldots\ldots\ldots, l_{je}^t\right]$b or j=1, 2, 3... N labels the locations of each crow *j* at cycle t, where l_j^t is the potential possible placement options for crow *j* in dimensions e.

If a crow *j* proclaims that it is interested in robbing another crow *i*, one of two things may happen:

According to Eq. (11), Crow *j* will locate Crow *i*'s food store and update Crow *i*'s location rather than really following Crow *i*.

$$l_j^{(t+1)} = l_j^{(t)} + r_j * ck_i^{(t)} * (n_i^{(t)} - l_j^{(t)}) \quad (11)$$

Where the journey distance, *ck* is shown. r_i was randomly selected from the range [0, 1].

i pursue the crow to see where its food is concealed after realizing it is an endangered species. The crow *i* in the scenario does irregular dancing to trick the crow *j*.

In reality, the two parts may be mixed numerically as seen below in Eq. (12);

$$l_j^{(t+1)} = \begin{cases} l_j^{(t)} + r_j * ck_i^{(t)} * \left(n_i^{(t)} - l_j^{(t)}\right), r_i \geq BP_j^t \\ Select\ a\ random\ position, \quad Otherwise \end{cases} \quad (12)$$

where, *j* and *i* are random integers between 0 and 1, BP_j^t is the probability that Crow *i* at Iteration *t* will be aware, and r_j and r_i are random numbers. The ability of the crow to seek is affected by the values of *ck*. Low values of *ck* will aid in local optimal while high values will aid in global search.

Every crow is assessed as the technique is running using a well-defined fitness function. The crows then relocate themselves according to where they are most comfortable. The viability of every new role is evaluated. Using Eq. (13), the crows' memory is enhanced.

$$n_j^{(t+1)} = \begin{cases} l_j^{(t+1)} jg\ c\left(l_j^{(t+1)}\right) & is\ better\ than\ c\left(l_j^{(t)}\right) \\ n_j^{(t)}, & Otherwise \end{cases} \quad (13)$$

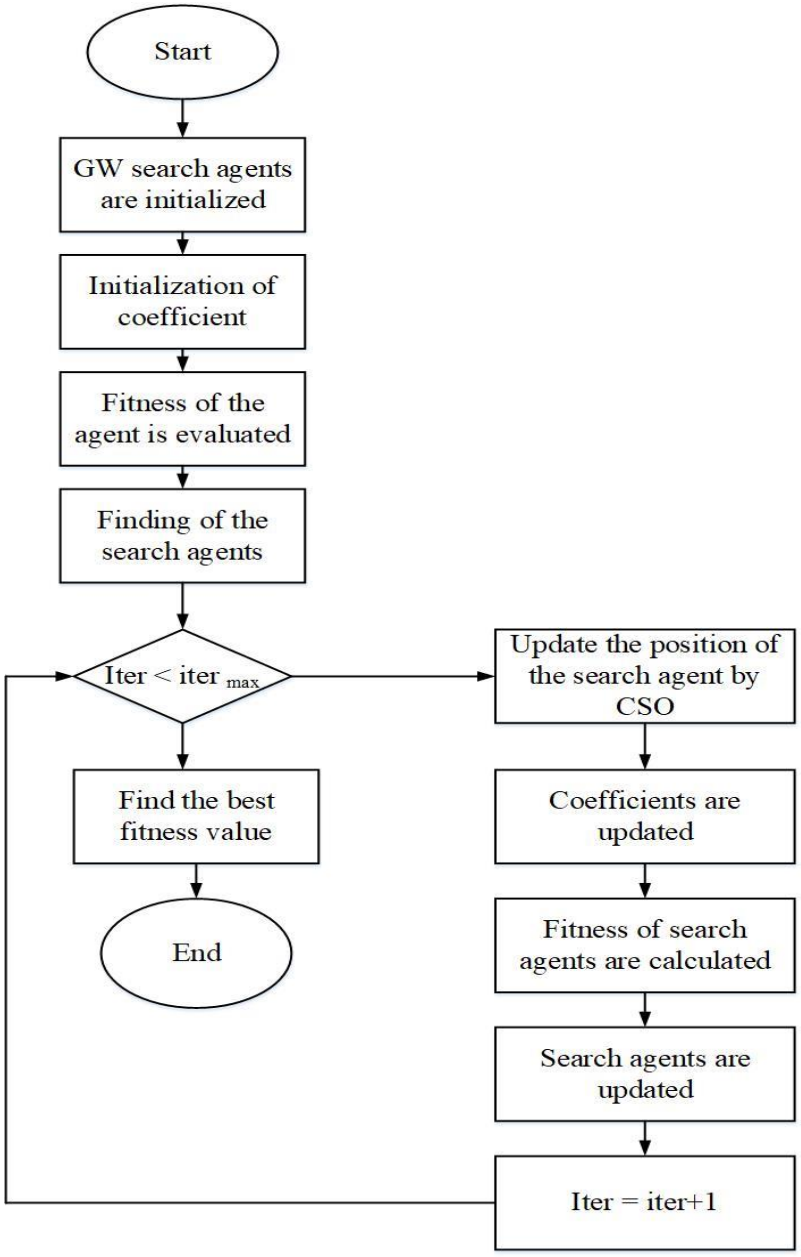

Fig. 3. Flow method of BGW-CS.

3) BGW-CS algorithm: To address the problem of immature integration, an improved research phase for BGWO is suggested in the current work. Through this method, BGWO and CSO algorithms are hybridized. In order to increase response accuracy, the created hybrid algorithm (BGW-CS) has a stronger propensity to pass over local optimal solutions. The flow methodology of BGW-CS is represented in Fig. 3. The equation of the hybrid BGW-CS is in Eq. (14),

$$A_{binary}(t+1) = \begin{cases} l_j^{(t+1)} jg\ c\left(l_j^{(t+1)}\right) & is\ better\ than\ c\left(l_j^{(t)}\right) \\ n_j^{(t)}, & Otherwise \end{cases} \quad (14)$$

D. Classification using Support Vector Machine

Apply the classification algorithm known as Support Vector Machine. Utilizing the multimodal medical data, train the SVM model with the pre-processed and chosen features to forecast diabetes outcomes because SVM can distinguish intricate correlations between features and is efficient in managing high-dimensional data, it is a popular choice. The SVM is used to group and stratify diverse classes of heterogeneous medical data, allowing for more precise predictions and individualised healthcare interventions. SVM uses its ability to construct hyperplanes that maximise the segregation of points of information in a multidimensional space of features to help categorise records of patients, outcomes of tests, and other important medical data. This categorization strategy is a crucial component of our model,

enhancing its accuracy and effectiveness in diabetes forecasting and providing healthcare practitioners with crucial information for better patient care and management approaches. The structural risk reduction idea is applied in SVM. Here, the learning machine's error rate is thought to be constrained by the interaction between the training error rate and the Vapnik Chervonenkis (VC) 1 dimension term. The hyper-plane that optimally distinguishes the data points for N training sets (X_i , Y_i) with labels, where $X_i \in$ Rn and $Y_i \in$ {–1, 1}, is illustrated in Eq. (15):

$$f(X_q) = \sum_{i=1}^{N} Y_i \propto_i K(X_q, X_i) + b \quad (15)$$

The sign of f (X_q) is used to calculate X_q in cases where the kernel functions are represented as K (.) and indicates the membership of the query sample. Making an ideal hyperplane is comparable to figuring out all nonzero _i, which stands for the bias b and the support vectors. The least amount of loss is anticipated while making a decision.

Using BGW-CSO to optimize the feature set and SVM to accurately forecast diabetes are the three main components of the comprehensive analysis methodology. Utilizing the advantages of each technique, this hybrid strategy seeks to improve the predictive model's overall performance.

V. Result and Discussion

The offered information from a research article describes a study that suggests a unique hybrid optimisation approach for enhancing the reliability and accuracy of diabetes prediction. To use multimodal medical data from the Pima Indian and Mendeley diabetes datasets, the model combines a Machine learning mechanism and an ensemble learning technique. To minimise the dimensions of the features, a hybrid technique known as Binary Grey Wolf-based Crow Search Optimisation (BGW-CSO) is utilized for selecting features. The selected features are then passed to a BGW-CSO-based Support Vector Machine (SVM) with enhanced hidden neurons. Utilising multiple performance criteria, the suggested model, BGW-CSO-SVM, is assessed and contrasted with existing methodologies. The findings show that the suggested methodology improves diabetes prediction accuracy, enabling early detection of at-risk patients and facilitating individualised patient management. According to the study, the hybrid optimisation model has the potential to significantly advance the science of diabetes prediction and help medical professionals make wise choices.

A. Evaluation of Performance Metrics

Accuracy, F1-score, precision, and recall were the four assessment measures used in the experiment to evaluate the models. These particular definitions of these parameters are given in Eq. (16) to Eq. (19):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (16)$$

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (18)$$

$$\text{F1score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall+precision}} \quad (19)$$

The total amount of data which were correctly classified as positive out of all the positive data is referred to as the TP. The number of data which were incorrectly classified as negative out from one of is referred to as the TN. FN, Is it common for the model to mistakenly classify positive data as negative when, in fact, they were positive in the dataset. FP, Is it common for the model to mistakenly classify data as positive when, in reality, they were negative in the dataset. Recall is the ratio of the strategy's correctly classified positive pieces of information to those correctly classified positive pieces of information in the data set. In terms of the total quantity of variables that were classified as positive, precision is the proportion of data that the algorithm correctly recognised as positive. The F1 score is the harmonic average of recall and accuracy. Table I compares the performance of various classification methods or models. The methods evaluated include Soft Voting Classifier, Random Forest, DMP_MI, Bootstrap Aggregation, and a Proposed BGW-CSO-SVM.

Among the efficacy metrics that are assessed are F1-score, recall, accuracy, and precision. Accuracy is a measure of how well the strategy projections as a whole were predicted. Precision is the proportion of accurately predicted positive occurrences among all positive forecasts, whereas recall measures the ability to recognise positive examples. A single statistic called the F1-score integrates accuracy and recall. The Presented BGW-CSO-SVM approach obtained the greatest accuracy, precision, recall, and F1-score, showing greater efficiency when compared with the other techniques, as can be shown in Table I.

In Fig. 4, the Proposed BGW-CSO-SVM demonstrates the highest accuracy with a score of 96.62%. Following closely behind is the Random Forest method with an accuracy of 94.1%. The Bootstrap Aggregation technique achieves an accuracy of 94.62%, while the DMP_MI method achieves 87.1% accuracy. The Soft Voting Classifier has the lowest accuracy among the methods, with a score of 79.08%.

TABLE I. Comparison of Performance Metrics with Proposed Model

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Soft Voting Classifier [26]	79.08	73.13	70	71.56
Random Forest [27]	94.1	97.6	94.3	95.9
DMP_MI[28]	87.1	80.6	85.4	83.0
Bootstrap Aggregation [29]	94.62	94.7	94.6	94.6
Proposed BGW-CSO-SVM	96.62	98.54	96.3	97.8

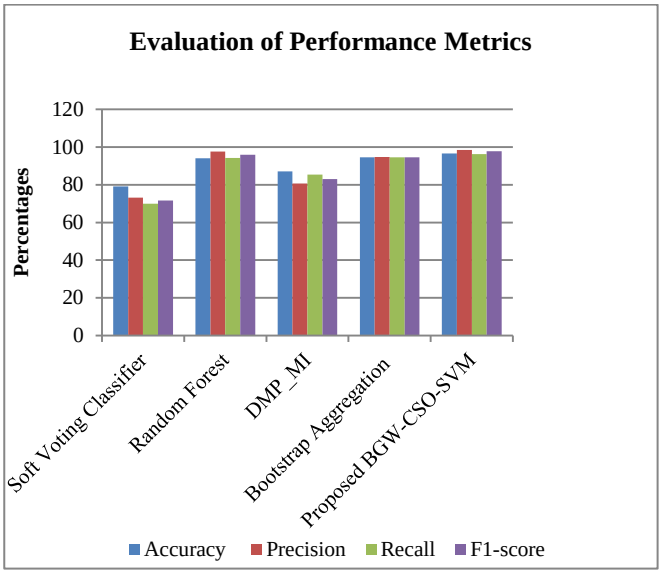

Fig. 4. Comparison graph of evaluation parameters.

In Table II, the proposed BGW-CSO-SVM model achieves the highest ROC value of 0.98, indicating strong discrimination capability between positive and negative instances in diabetes prediction. The Soft Voting method also demonstrates a high ROC value of 0.96, followed by LightGBM with a value of 0.95. XGBoost and Random Forest have ROC values of 0.93 and 0.94, respectively. AdaBoost shows a slightly lower ROC value of 0.92. Fig. 5 depicts the performance assessment of ROC curve. These results suggest that the proposed BGW-CSO-SVM model outperformed the other methods in terms of ROC value, indicating its potential for accurate prediction and effective management of diabetes.

The effectiveness of a classifier is often assessed using the ROC values in binary categorization tasks. Greater numbers denote greater performance. It shows a model's capacity to discriminate among both positive and negative examples.

A. Dataset Comparison

In Table III, the efficacy of the suggested BGW-CSO-SVM model in comparison to the approach presented by Taz, Islam, and Mahmud [30] is evaluated using two datasets: the Mendeley Diabetes Dataset and the PID Dataset. Fig. 6 depicts the proposed system dataset compared with existing approach. The evaluation metrics used are accuracy, precision, recall, and F1-score. The proposed BGW-CSO-SVM model consistently achieves higher accuracy, precision, recall, and F1-score on both the Mendeley Diabetes Dataset and the PID Dataset compared to the approach.

Table IV and Fig. 7 presents the results of a diabetes prediction study using different methods and their respective Root Mean Square Error (RMSE) values. The methods evaluated are Vanila-LSTM, BI-LSTM, and the Proposed Model.

TABLE II. COMPARISON OF PROPOSED SYSTEM ROC VALUE WITH EXISTING APPROACH

Methods	ROC value
Light GBM [30]	0.95
XGBoost [30]	0.93
AdaBoost [30]	0.92
Random Forest [30]	0.94
Soft Voting [30]	0.96
Proposed BGW-CSO-SVM	0.98

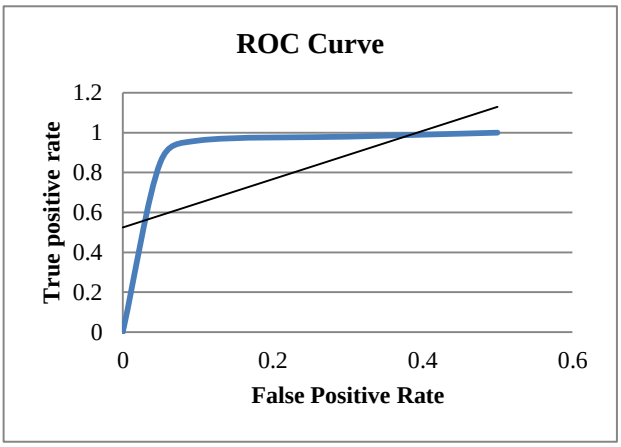

Fig. 5. ROC curve for the proposed model.

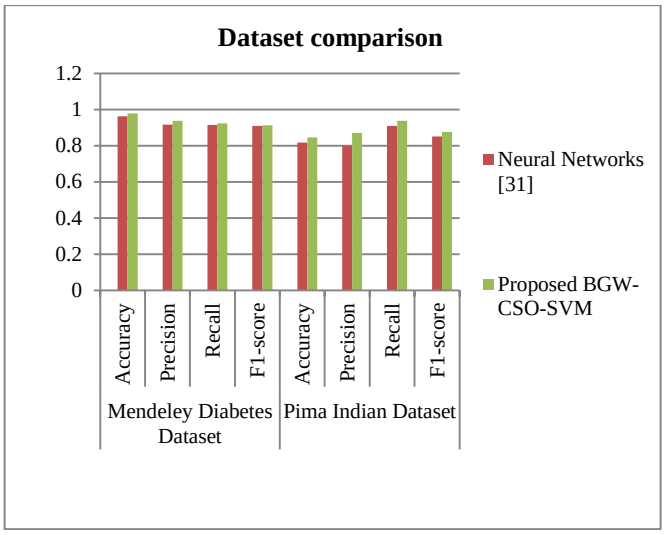

Fig. 6. Dataset comparison.

TABLE III. DATASET COMPARISON MDD AND PID

Methods	Mendeley Diabetes Dataset				Pima Indian Dataset			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Neural Networks [30]	96.24%	91.65%	91.46%	90.96%	81.82%	80.00%	90.91%	85.11%
Proposed BGW-CSO-SVM	97.9%	93.8%	92.36%	91.4%	84.63%	87.1%	93.8%	87.6%

TABLE IV. ERROR RATE COMPARISON

Method	RMSE
Vanila-LSTM [31]	15.43
BI-LSTM [31]	15.22
Proposed SVM	14.58

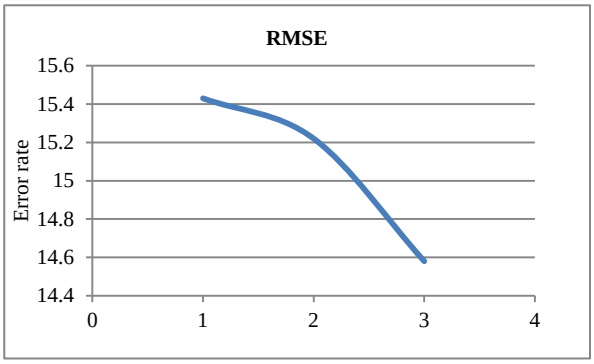

Fig. 7. Error rate graph.

B. Discussion

The study employs a feature selection technique, minimizes feature size, and ensures error-free classifications to effectively solve the difficulty of high-dimensional feature space in diabetes prediction. The research provides a novel methodology for feature selection, improving the model's capacity for generalization. It does this by introducing the revolutionary Binary Grey Wolf-based Crow Search Optimization (BGW-CSO) method, which combines Binary Grey Wolf Optimization (BGWO) with Crow Search Optimization (CSO). By optimizing the number of hidden neurons in the Support Vector Machine (SVM) through the use of BGW-CSO, the study improves diabetes prediction models overall and enhances the effectiveness of conventional SVMs. The suggested BGW-CSO-SVM model consistently outperforms other existing techniques, demonstrating superior accuracy, precision, recall, and F1-score through rigorous assessments on the Mendeley Diabetes Dataset and Pima Indian Dataset. The study's objective of creating a reliable diabetes prediction model that can be adjusted to fit a variety of datasets is perfectly aligned with this performance. Most importantly, the hybrid optimization model BGW-CSO-SVM greatly improves diabetes prediction accuracy, allowing for early at-risk individual identification, timely interventions, and individualized patient management. The results of the study highlight how well the objective of increasing the accuracy and consistency of diabetes prediction was met, and they may have ramifications for better patient outcomes in medical settings.

VI. CONCLUSION AND FUTURE WORK

The study addresses the pressing global health challenge of accurate and timely diabetes diagnosis. It introduces a novel hybrid optimization strategy that seamlessly integrates machine learning techniques to enhance the precision of diabetes prediction. This advancement is achieved by harnessing the potential of multimodal medical data, electronic health records, and advancements in data analytics. To ensure precise classification, two distinct datasets from the Pima Indian diabetes databases are used, alongside a rigorous feature selection method. The introduction of the BGW-CSO approach, a fusion of BGWO and CSO, bolsters feature selection capabilities. This innovation not only addresses the challenges posed by high-dimensional feature spaces but also significantly improves the system's ability to generalize. The performance of conventional Support Vector Machines (SVMs) benefits greatly from optimizing SVM techniques using the newly devised BGW-CSO approach. This proposed strategy, referred to as BGW-CSO-SVM, demonstrates substantial enhancements in diabetes prediction accuracy, as evidenced by a comprehensive examination using various performance metrics and comparisons with existing methodologies. This breakthrough enables the rapid identification of individuals at risk, paving the way for personalized and effective treatment interventions. In order to ensure the applicability of our methodology across a wide range of healthcare contexts, future research endeavours should prioritize testing its usefulness across varied datasets and demographics. It is important to prioritize creating solutions that are easy to use, accessible, and readily embraced by medical professionals. By facilitating a smooth transition into clinical settings, this strategy hopes to improve patient care and diabetes diagnosis on a larger scale. The research can demonstrate the robustness and generalizability of the suggested technique, increasing the likelihood of its widespread adoption and beneficial effects on healthcare practices. This can be achieved by expanding the validation to include a broad variety of datasets and demographic differences.

REFERENCES

[1] L. Garcia-Molina, A.-M. Lewis-Mikhael, B. Riquelme-Gallego, N. Cano-Ibanez, M.-J. Oliveras-Lopez, and A. Bueno-Cavanillas, “Improving type 2 diabetes mellitus glycaemic control through lifestyle modification implementing diet intervention: a systematic review and meta-analysis,” Eur. J. Nutr., vol. 59, no. 4, pp. 1313–1328, 2020, doi: https://doi.org/10.1007/s00394-019-02147-6.

[2] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” Front. Genet., vol. 9, p. 515, 2018, doi: https://doi.org/10.3389/fgene.2018.00515.

[3] T. Latchoumi, J. Dayanika, and G. Archana, “A comparative study of machine learning algorithms using quick-witted diabetic prevention,” Ann. Romanian Soc. Cell Biol., pp. 4249–4259, 2021.

[4] J. J. Wong et al., “Depression in context: Important considerations for youth with type 1 vs type 2 diabetes,” Pediatr. Diabetes, vol. 21, no. 1, pp. 135–142, 2020, doi: https://doi.org/10.1111/pedi.12939.

[5] G.-M. Huang, K.-Y. Huang, T.-Y. Lee, and J. T.-Y. Weng, “An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients,” in BMC bioinformatics, BioMed Central, 2015, pp. 1–10. doi: https://doi.org/10.1186/1471-2105-16-S1-S5.

[6] J. Zhu, Q. Xie, and K. Zheng, “An improved early detection method of type-2 diabetes mellitus using multiple classifier system,” Inf. Sci., vol. 292, pp. 1–14, 2015, doi: https://doi.org/10.1016/j.ins.2014.08.056.

[7] L. A. Sleeper et al., “Evaluation of Kawasaki disease risk-scoring systems for intravenous immunoglobulin resistance,” J. Pediatr., vol. 158, no. 5, pp. 831–835, 2011, doi: https://doi.org/10.1016/j.jpeds.2010.10.031.

[8] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, “Current techniques for diabetes prediction: review and case study,” Appl. Sci., vol. 9, no. 21, p. 4604, 2019, doi: https://doi.org/10.3390/app9214604.

[9] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017, doi: https://doi.org/10.1016/j.csbj.2016.12.005.

[10] A. J. Vickers, A. M. Cronin, E. B. Elkin, and M. Gonen, "Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers," BMC Med. Inform. Decis. Mak., vol. 8, pp. 1–17, 2008, doi: https://doi.org/10.1186/1472-6947-8-53.

[11] C. Chopra, S. Sinha, S. Jaroli, A. Shukla, and S. Maheshwari, "Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients," in proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, 2017, pp. 18–23. doi: https://doi.org/10.1145/3155077.3155081.

[12] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework," J. Biomed. Inform., vol. 59, pp. 185–200, 2016, doi: https://doi.org/10.1016/j.jbi.2015.12.001.

[13] P. P. Debata and P. Mohapatra, "Diagnosis of diabetes in pregnant woman using a Chaotic-Jaya hybridized extreme learning machine model," J. Integr. Bioinforma., vol. 18, no. 1, pp. 81–99, 2020, doi: https://doi.org/10.1515/jib-2019-0097.

[14] ISTI, "Exploring Diabetes Epidemic in India | India Science, Technology & Innovation - ISTI Portal." Accessed: Jun. 23, 2023. [Online]. Available: https://www.indiascienceandtechnology.gov.in/featured-science/exploring-diabetes-epidemic-india

[15] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Inform. Med. Unlocked, vol. 16, p. 100203, 2019, doi: https://doi.org/10.1016/j.imu.2019.100203.

[16] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Inform. Med. Unlocked, vol. 17, p. 100179, 2019, doi: https://doi.org/10.1016/j.imu.2019.100179.

[17] S. Cui, D. Wang, Y. Wang, P.-W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," Comput. Methods Programs Biomed., vol. 166, pp. 123–135, 2018, doi: https://doi.org/10.1016/j.cmpb.2018.10.012.

[18] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," Int. J. Cogn. Comput. Eng., vol. 2, pp. 229–241, 2021, doi: https://doi.org/10.1016/j.ijcce.2021.12.001.

[19] G. Cappon, F. Prendin, A. Facchinetti, G. Sparacino, and S. D. Favero, "Individualized Models for Glucose Prediction in Type 1 Diabetes: Comparing Black-box Approaches To a Physiological White-box One," IEEE Trans. Biomed. Eng., pp. 1–11, 2023, doi: 10.1109/TBME.2023.3276193.

[20] J. Xie and Q. Wang, "Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison With Classical Time-Series Models," IEEE Trans. Biomed. Eng., vol. PP, Feb. 2020, doi: 10.1109/TBME.2020.2975959.

[21] M. Alirezaei, S. T. A. Niaki, and S. A. A. Niaki, "A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines," Expert Syst. Appl., vol. 127, pp. 47–57, 2019, doi: https://doi.org/10.1016/j.eswa.2019.02.037.

[22] R. D. H. Devi, A. Bai, and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," Obes. Med., vol. 17, p. 100152, 2020, doi: https://doi.org/10.1016/j.obmed.2019.100152.

[23] K. S. Prasad, N. C. S. Reddy, and B. Puneeth, "A framework for diagnosing kidney disease in diabetes patients using classification algorithms," SN Comput. Sci., vol. 1, no. 2, p. 101, 2020, doi: https://doi.org/10.1007/s42979-020-0096-7.

[24] P. Nuankaew, S. Chaising, and P. Temdee, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction," IEEE Access, vol. 9, pp. 137015–137028, 2021, doi: 10.1109/ACCESS.2021.3117269.

[25] A. Askarzadeh, "A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm," Comput. Struct., vol. 169, pp. 1–12, 2016.

[26] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," Int. J. Cogn. Comput. Eng., vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[27] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," Procedia Comput. Sci., vol. 167, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[28] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," IEEE Access, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[29] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," Sensors, vol. 22, no. 14, p. 5247, Jul. 2022, doi: 10.3390/s22145247.

[30] N. H. Taz, A. Islam, and I. Mahmud, "A Comparative Analysis of Ensemble Based Machine Learning Techniques for Diabetes Identification," in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), DHAKA, Bangladesh: IEEE, Jan. 2021, pp. 1–6. doi: 10.1109/ICREST51555.2021.9331036.

[31] H. Butt, I. Khosa, and M. A. Iftikhar, "Feature Transformation for Efficient Blood Glucose Prediction in Type 1 Diabetes Mellitus Patients," Diagnostics, vol. 13, no. 3, p. 340, Jan. 2023, doi: 10.3390/diagnostics13030340.