



RESEARCH ARTICLE

Hyperparameter tuning of diabetes prediction using machine learning algorithm with pelican optimization algorithm

T. Ramyaveni*, V. Maniraj

Abstract

Machine learning algorithms are employed in public health to forecast or diagnose chronic epidemiological illnesses like diabetes, which have global rates of transmission and infection. Machine learning technology may be applied to diagnostic, prognostic, and evaluation methods for a number of illnesses, including diabetes. This work presents a novel approach based on a novel metaheuristic optimization algorithm to improve diabetes categorization. About 738 records were included in the final analysis of the main data, which was acquired in 2013 in accordance with the security protocols specified in the Declaration of Helsinki. This approach suggests a novel feature selection technique based on douche optimization technique (DBERDTO) and the dynamic Al-Biruni earth radius. A random forest classifier was used to categorize the chosen features, and the suggested DBERDTO was utilized to optimize the parameters. In this work, we investigate hyperparameter tuning for improved diabetes case prediction using the pelican optimization algorithm (POA) in conjunction with the XGBoost machine learning technique. To prove the effectiveness and superiority of the suggested approach, it is tested against the most recent machine learning models and optimization techniques. The method's overall accuracy for classifying diabetes was 99.65%. These test results attest to the suggested method's superiority over alternative categorization and optimization techniques.

Keywords: Machine learning algorithms, Diabetes mellitus, Helsinki declaration, Al-Biruni earth radius, Dipper-throated optimization algorithm, Pelican optimization algorithm.

Introduction

Diabetes mellitus (DM) is a metabolic condition characterized by hyperglycemia brought on by aberrant insulin production, faulty insulin action, or both. Diabetes-related chronic hyperglycemia can cause long-term harm, malfunction, and failure to organs such as the kidneys, heart, eyes, blood vessels, and nerves (Alhussan, A. A., *et al.*, 2023). Based on the etiology and clinical circumstances, diabetes mellitus

(DM) is classified into three subtypes: type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM), and gestational diabetes mellitus. A cell-mediated autoimmune response in type 1 diabetes frequently leads to the death of pancreatic beta cells, which leaves the patient completely insulin-deficient. Insulin resistance and moderate insulin insufficiency are the main causes of type 2 diabetes. Different levels of glucose intolerance are a hallmark of gestational diabetes, sometimes referred to as gestational diabetes (El-Kenawy, E. S. M., *et al.*, 2023). Type 2 diabetes affects the majority of diabetics. People of all ages can be affected by type 2 diabetes, although those over 40 are more likely to have it. These symptoms could not show up for years, and many patients receive an unintentional diagnosis after undergoing therapy for unrelated issues. Even if individuals with type 2 diabetes do not require insulin, insulin therapy can be necessary if diet and oral antidiabetic medication treatment are insufficient to manage hyperglycemia (Shazly, K.; Khodadadi, N.(2023)) Type 2 diabetes has a wide range of intricate causes. Not all of the variables that affect an individual's risk of contracting the disease are direct causes. These variables might be behavioral (food, smoking, obesity, lack of exercise), genetic, or demographic (such as age). The International Diabetes Federation (IDF) reported that in 2019,

Department of Computer Science, A.V.V.M Sri Pushpam College (Autonomous), Poondi, Thanjavur (Affiliated to Bharathidasan University), Tiruchirapalli, Tamil Nadu, India.

***Corresponding Author:** T. Ramyaveni, Department of Computer Science, A.V.V.M Sri Pushpam College (Autonomous), Poondi, Thanjavur (Affiliated to Bharathidasan University), Tiruchirapalli, Tamil Nadu, India., E-Mail: veniramyas@gmail.com

How to cite this article: Ramyaveni, T., Maniraj, V. (2024). Hyperparameter tuning of diabetes prediction using machine learning algorithm with pelican optimization algorithm. The Scientific Temper, **15**(3):2616-2622.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.29

Source of support: Nil

Conflict of interest: None.

over 460 million individuals globally received a diabetes diagnosis; by 2030 and 2045, that figure is expected to rise to 578 million. It is anticipated that this figure will increase to 700 million. Type 2 diabetes mellitus (T2DM) is becoming more and more common and is one of the most alarming diseases.

Diabetes type 2 has many different and intricate causes. Though not all of them are direct causes, a number of factors might raise or lower the likelihood of contracting the illness. Genetic, demographic, or behavioral variables (food, smoking, obesity, lack of exercise) might be among them. Because behavioral risk factors are adaptable and may be enhanced, they are also known as “modifiable” risk factors (Oubelaid, A., *et al.*, 2022). Over 460 million individuals worldwide were diagnosed with diabetes in 2019, and by 2030 and 2045, that figure is expected to rise to 578 million, according to research published by the International Diabetes Federation (IDF). Up to 700 million people are predicted to live in this country. One of the most alarming trends in healthcare today is the rising prevalence of type 2 diabetic mellitus (T2DM). Diabetes modeling research has made use of a number of machine learning approaches, such as decision trees (DT), k closest neighbor (KNN), support vector machine (SVM), and artificial neural network (ANN) (Almutairi, E. S., & Abbod, M. F. (2023)), While studying diabetes prevalence patterns and utilizing population-specific risk variables to forecast future burdens are crucial, machine learning categorization approaches have not received much attention.

The study shows that the incidence of NAFLD/NASH is elevated in North Karnataka, which is 31.23%. The occurrence of diabetic mellitus and obesity with high BMI is a major risk factor for the growth of NASH. Fibroscan which is an easy and non-invasive test for assessing the determination of the liver along with liver function tests, can be used to foretell significant fibrosis of the liver that is NASH. This is also the side effect of an increase in diabetics (Desai, G. S., *et al.*, 2023).

When it comes to examining medical datasets, such as those pertaining to diabetes, feature selection is essential. The objective is to identify a subset of attributes that significantly influence categorization or prediction processes. The issue lies in the fact that overfitting and poor classification performance might result from high-dimensional data sets that contain redundant or unimportant characteristics. Researchers looked on selecting features in diabetes datasets using metaheuristic optimization techniques in order to solve this problem. Designed to effectively investigate a wide range of possible solutions, these algorithms were inspired by real-life circumstances and troubleshooting techniques. These techniques are meant to increase the dimensionality reduction, readability, and accuracy of diabetes classification models (Chou, C. Y., *et al.*, 2023).

The chronic metabolic condition known as diabetes is posing a serious threat to health systems worldwide. Machine learning algorithms, particularly XGBoost, have been utilized to provide practical outcomes in predictive analytics in the field of medical diagnostics. Important configuration choices that impact algorithm performance are called hyperparameters. As such, the success of these models depends on precise adjustment. Finding effective and dependable diabetes prediction models requires combining XGBoost with contemporary hyperparameter tuning techniques (Katarya, R., & Polipireddy, S. (2020)).

XGBoost is a well-known, efficient and successful ensemble learning technique in a variety of industries, including medical. Predictive modeling is popular because it can handle intricate relationships in data sets. To get good results. Nevertheless, substantial and precise hyperparameter adjustment is needed over time. The principal inputs consist of:

- This work aims to explore hyperparameter optimization techniques that might enhance
- XGBoost’s diabetic prediction performance. To look at the most significant possible risk factors of type 2 diabetes, both the PCA and IG techniques were used. The dataset’s dimensionality was also decreased by doing a multivariable feature analysis.
- To investigate the key possible risk factors for type 2 diabetes, we used both PCA and IG techniques. To further minimize the dataset’s dimensionality, a multivariable feature analysis was carried out.
- In order to make sure that the model generalizes to new data in an efficient manner, the internal parameters of the model must be adjusted using hyperparameter optimization.

Related Work

In the last ten years, machine learning has become increasingly popular for predicting a patient’s risk of acquiring diabetes as well as for identifying risk factors for the disease. ML is utilized not only to forecast the onset of diabetes but also its associated consequences, including retinopathy, neuropathy, and kidney failure (Febrian, M. E., *et al.*, 2023). Researchers employed psychological information to forecast depression in diabetes patients as part of a study on the disease’s consequences (Ellulu, M. S., & Samouda, H. (2022)). Many researchers obtain data from secondary sources, such as electronic duty records (EMRs), as opposed to using surveys to gather data EMR containing lipid profile data, such as blood glucose levels, triglyceride (TG) levels, etc. for diabetes prediction, as well as clinical and non-clinical data that are ideally suited for predictive analytics. Numerous exist (Eid, M. M., *et al.*, 2022). Conversely, several researchers have assessed the popular Pima Indian Diabetes Database (PIDD), but their attempts have not yielded sufficient levels of predicted accuracy. In

the PIDD example, the accuracy obtained by (Mahmood, A., *et al.*, 2022) was 75%, (Mahmood, A., *et al.*, 2022) was 80%, and (Forbus, J. J., & Berleant, D. (2022)) was 77%. In addition, most underdeveloped nations—including Bangladesh—lack pertinent and trustworthy marker data for research on diabetes prediction. This emphasizes the requirement for more accurate diabetes datasets with a focus on the local area. As a result, this study is very different from other studies since it used primary sources to gather data in Bangladesh and other nations with high diabetes incidence.

Furthermore, a variety of variables have been employed in several studies to predict diabetes. For instance, body mass index (BMI), exercise, high blood pressure, age, gender, and family history are often employed characteristics for machine learning-based diabetes prediction (Abdelhamid, A. A., & Alotaibi, S. R. (2022)). When categorizing populations at high risk for diabetes, the World Health Organization (WHO) takes into account both non-clinical (such as demographic characteristics) and clinical (such as blood glucose levels) criteria (Khafaga, D. (2022)). Moreover, there is mounting evidence in nations that are experiencing significant economic growth that the number of individuals with diabetes is inextricably connected to a rise in the level of living. Therefore, as part of integrating several features to predict diabetes, we chose all clinical, non-clinical, and dietary components in this investigation. In addition, this study adds to the body of knowledge by examining dietary components that are important in the development of diabetes, especially in South Asia.

In this article, the author has implemented PPML techniques, particularly federated learning, which hold vast areas in revolutionizing healthcare data sharing by allowing mutual model development without revealing raw patient data. Privacy-preserving techniques must be adjustable to these variations to ensure data security and regulatory compliance. The equal distribution of measures across different data modalities fosters data fairness and equity, ensuring that all healthcare data types are effectively represented in research, eventually leading to a more comprehensive understanding of patients' health (Ahamed, S. K., *et al.*, 2023).

Methodology

Dataset

About 738 individuals from different Bangladeshi urban and rural locations provided data for the study using questionnaires. Three main components made up the survey questions: non-clinical, clinical, and demographic characteristics. We first sought input from two diabetes specialists and a public health researcher before developing the questionnaire and choosing question items based on a study of prior research. Initially, we chose 18 characteristics. Given that the questionnaire contained clinical data (blood

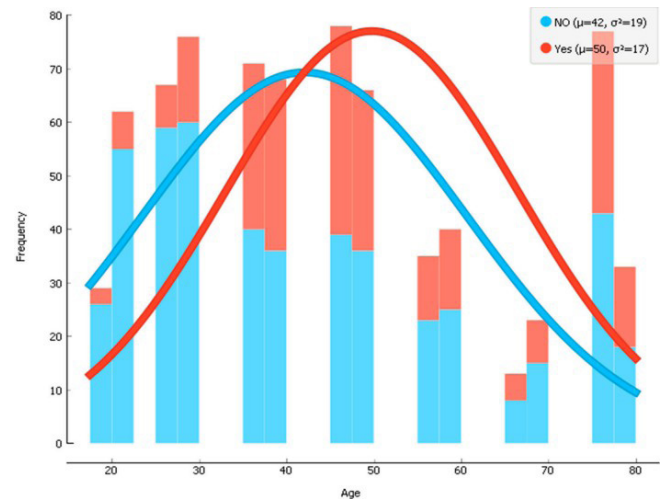


Figure 1: The histogram of the age of participants data preprocessing

pressure, blood sugar levels, etc.), all participants gave their previous agreement before beginning the study, and participation was entirely voluntary. The study's goal, expected outcomes, and societal advantages were all stated. Following manual review, we were able to extract data on both healthy individuals and diabetes patients.

There were 482 non-diabetic volunteers and 256 diabetes patients (or around 34.6% of the 738 participants). Participants ranged in age from 19 to 78 years old, with an average age of 44. We divided the data set in half before training, using 80% of the data for the training set and the remaining 20% for the test set. This ratio was 80:20. The model is then fed the training data in order to use guided machine learning techniques. Figure 1 displays the age histograms for persons with diabetes (mean age = 50) and those without diabetes (mean age = 42).

Preprocessing transforms the data so that you may create machine learning models that are more accurate. Preprocessing includes a number of operations, including normalizing the data, eliminating records with missing values, and rejecting outliers. To prepare the dataset for fitting, we utilize MinMaxScaler to normalize the data values. In 482 samples in the dataset were found to be free of diabetes, while 256 samples were verified to have the disease. After that, the data is input into the model by being separated into training and test sets. The provided dataset is split 80:20 for training and testing, with the former serving as the main objective.

Feature Selection

An essential step in the diabetes classification process is determining which characteristics are most appropriate and contribute the most to classification accuracy. Feature selection eliminates irrelevant or redundant information, reducing dimensionality and increasing classification

accuracy. As a result, the model runs more quickly and effectively in real-time scenarios and becomes simpler. The process of determining which characteristics are most helpful by assessing them according to a set of standards, such as how well they correlate with a target variable and how well they can differentiate across classes, is known as feature selection. There are several approaches to feature selection, including wrappers, filters, and embedding techniques.

Filtering techniques utilize statistical tests or correlation coefficients to assess each feature's significance outside of a classification model. On the other hand, wrapping strategies add and delete features iteratively based on the relative significance determined by the classification model. The embedded technique integrates the feature selection and classification model training processes into one step, where feature weights are learned directly from the data. In order to improve the accuracy and interpretability of the diabetes classification model, feature selection is crucial. Numerous investigations have demonstrated that feature selection techniques may greatly reduce a dataset's dimensionality without compromising classification accuracy.

Researchers have discovered that feature selection procedures outperform utilizing all characteristics for classification; the greatest results are obtained when feature selection is based on mutual information. Additionally, according to the experts, feature selection may aid in determining the key characteristics linked to GDM, thus enhancing diagnostic instruments. By emphasizing the most crucial components and simplifying the data set, feature selection helps make analysis and interpretation easier to handle. To help the condition, you may also select a unique diagnostic. The particular diagnosis that has the most data may be further refined and is a straightforward and uncomplicated diagnostic technique. Excellent, thorough number collections, illustrations of number collections related to diabetes, and effective resources for creating enhanced arithmetic procedures in the Genki style. The local region is the best and the best results are coming closer when it comes to natural phenomena or human activities. This is the most efficient approach to compute the effective search space. For the specific selection of diabetes categorization, metacyclic calculation techniques (such as the integrated genetic calculation method, particle swarm improvement, cloud swarm improvement, model fire extinguishing, and artificial bee colony) are frequently utilized. These algorithms provide fitness functions based on information retrieval, classification accuracy, and other criteria that you may use to search for collections of high-level characteristics. In addition to being faster, metaheuristic optimization can get around the drawbacks of other feature selection techniques, including wrappers, filters, and embedding techniques. When used to choose features for diabetes classification, metaheuristic

optimization worked well and might result in the creation of more reliable and accurate models.

The binary numbers 0 and 1 make up the tiny search space used in the feature selection problem to determine the significance of a particular feature. We suggest a binary version of the DBERD TO approach to assist with better adaptation to the feature selection procedure. The original algorithm's continuous values should be treated as binary [0, 1] values. The following formula provides the *Sigmoid* Sigmoid function, which is utilized to convert values to binary.

$$S^{t+1} = \begin{cases} 1 & \text{if } Sigmoid(S_{best}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$Sigmoid(S_{Best}) = \frac{1}{1 + e^{-10(S_{Best} - 0.5)}} \quad (2)$$

In the T stage of the iterative process, S_{best} is the best option. Scaling continuous values to the discrete range [0,1] is the primary goal of the sigmoid function.

Hyperparameter Tuning

Pelican optimization algorithm and XGBoost

The pelican optimization method (POA) is a metaheuristic method that may be used to solve optimization issues, including hyperparameter tweaking. The POA method randomly initializes a set of potential solutions, symbolized by the pelican, at the start of the search area. Every pelican represents a potential set of XGBoost algorithm hyperparameters. The technology iteratively modifies the pelican's posture in three fundamental processes (food finder, mushy food finder, bubble net feeding cycle).

All pelicans change positions to navigate across the navigation space during the feeding navigation phase. This inquiry is required to prevent being mired in a solution and contributes to a more thorough examination of the search process overall. When pelicans congregate around food, they select the most fitting object to represent it. Both solution development and convergence are helpful at this point. Eventually, as they dine on the foam netting, the pelicans circle around the bait to form a group.

During the bubble net feeding stage, we used the following equation to update the locations of the pelicans:

$$X' = X_{best} - A * D \quad (3)$$

D is the distance between the pelican's current location and the bait position, and X' is the most recent position of the bird, which is updated continually. The constant b determines the pattern's form. An integer l, between 0 and 1, that is created at random to indicate the current best solution location.

When the pelican is in the bubble net feeding phase, its position is updated using the formula below: To further alter the search agent's position during the bubble net configuration stage, utilize the formula below:

$$X' = D * e^{(b * l)} * COS(2\pi * l) + X_{best} \quad (4)$$

X' prime represents the updated location, the current optimal solution position is represented by X_best , b is the shape control constant, l is a random number between 0 and 1, and D is the distance between the position and the feed position.

Equations (3) and (4) are used to update the pelican locations so that the POA method can explore hyperparameter space and converge on an optimal set of values customized for the XGBoost approach.

Results and Discussion

All machine learning algorithms rely on a number of parameters, including precision (the model’s ability to function correctly) and accuracy (the ratio of accurate predictions to total predictions). To gauge the model’s performance, I employed a confusion matrix. Affirmative forecasts), specificity, recall (the ratio of all true positive predictions to all positives for the actual class), and F1 score (only the weighted average of recall and accuracy). The final row of Table 1 displays the formula for this parameter. The confusion matrix is helpful in determining when a classification model becomes confused during prediction as it computes true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). We used the confusion matrix to evaluate the performance of the classification model/algorithm and then plotted the receiver operating characteristic (ROC) curve to calculate the performance of each model under various thresholds (starting at 0). After step 1), determine the area under the ROC curve, or AUC.

Performance Metrics

In this paper, we define this topic as a machine learning binary classification problem. The accuracy score obtained from the test data is, thus, the main performance metric we use. The F1 score (F-measure), recall, and accuracy of the proposed model were also calculated.

Accuracy

For assessing the effectiveness of classification algorithms, this is the most often used metric. One possible definition for it would be the ratio of accurate forecasts to total forecasts. The confusion matrix and the following formula may be combined to compute it in an easy method.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

Table 1: Confusion matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Precision

Our machine learning model’s ability to retrieve accurate documents may be measured by calculating the total number of documents it can retrieve. We can calculate it rapidly with the confusion matrix and the following formula.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Recall

One way to characterize recall would be the quantity of false positives generated by our machine learning model. Using the confusion matrix and the following formula, it is simple to calculate.

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

F1-Score

When accuracy and memory contribute proportionately equally, the F1 score is attained.

$$F1 - score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{11}$$

Experimental Results

A ROC curve is created when all of the sample points are joined together to form a line. The TPR, or the proportion of right judgments, increases with the line’s proximity to the top. Put differently, as Figure 2 illustrates, the higher the performance, the greater the area under the ROC curve (AUC).

In order to predict diabetes, this study uses the XGBoost algorithm in conjunction with pelican optimization. The obtained performance metrics are displayed in Table 2.

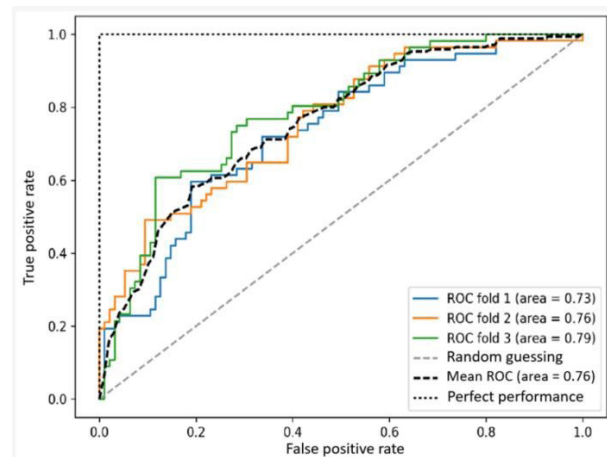


Figure 2: ROC curve

Table 2: Performances of classification methods from LIDC dataset

Measures	WOA	BOA	MFO	XGBoost-POA
Accuracy	88.36	93.98	93.82	94.69
Sensitivity	85.55	94.98	94.88	96.61
Specificity	91.98	95.68	95.82	98.73
Recall	92.51	97.98	96.82	99.63

Table 3: Comparison of the current study with the state of the art

Method	Accuracy
[10]	92.38
[12]	91.36
[15]	85.28
[14]	95.92
[17]	97.43
XGBoost-POA	99.65

This section explains classification analysis and demonstrates how to apply it to various tasks. Compared to the WOA, BOA, and MFO models, the approach suggested in Table 2 increases accuracy by 88.36, 93.98, and 93.82%. The POA model that was used performs 94.69% better than WOA, BOA, and MFO approaches while taking sensitivity into account. The POA model that has been implemented exhibits a 99.63% performance gain over WOA, BOA, and MFO approaches while taking recall rate into account. A thorough investigation demonstrates the superiority of the XGBoost-POA architecture.

To give a good performance evaluation, we use the same dataset to compare our optimization model with earlier research. This is the outcome of the five methodologies compared in this paper being categorized. The accuracy indicator comparison results are shown in Table 3. The comparison demonstrates the higher accuracy of the upgraded diabetes prediction categorization algorithm. [13] At 97.43% accuracy, the deep neural network demonstrated strong performance as well. In conclusion, even with a small training data set, the XGBoost-POA optimization approach may yield good results.

Conclusion

This work offers a novel approach based on metaheuristic optimization to improve diabetes categorization. We applied the dynamic Al-Biruni earth radius and throat optimization (DBERD TO) technique to develop a novel feature selection algorithm. Prior to applying it to the dataset, the suggested DBERD TO is used to improve the XGBoost parameters using the Pelican optimization method classifier once the features have been chosen. This study's primary goal was to investigate the differences between clinical and non-clinical characteristics in predicting diabetes in Bangladesh, a nation whose population, food habits, and standard of living are changing quickly. To sum up, this work effectively demonstrates the efficacy of the proposed method for hyperparameter optimization in the context of diabetes prediction, which integrates the XGBoost machine learning algorithm with the POA. Our dataset uses a dataset of 768 patient records in order to extract insightful information. The improved XGBoost model performs exceptionally well, achieving a peak accuracy of 99.65%. In the future, this

research will improve the accuracy of disease prediction and lead to more personalized medical services and early intervention.

Acknowledgments

The authors would like to thank A.V.V.M Sri Pushpam College (Autonomous) Poondi, Thanjavur, which is affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu for helpful supporting this work in an effective manner.

References

- Abdelhamid, A. A., & Alotaibi, S. R. (2022). Robust Prediction of the Bandwidth of Metamaterial Antenna Using Deep Learning. *Computers, Materials & Continua*, 72(2).
- Ahamed, S. K., Nishant, N., Selvaraj, A., Gandhewar, N., Srithar, A., & Baseer, K. K. (2023). Investigating privacy-preserving machine learning for healthcare data sharing through federated learning. *The Scientific Temper*, 14(04), 1308-1315.
- Alhussan, A. A., Abdelhamid, A. A., Towfek, S. K., Ibrahim, A., Eid, M. M., Khafaga, D. S., & Saraya, M. S. (2023). Classification of diabetes using feature selection and hybrid Al-Biruni earth radius and dipper throated optimization. *Diagnostics*, 13(12), 2038.
- Almutairi, E. S., & Abbod, M. F. (2023). Machine learning methods for Diabetes prevalence classification in Saudi Arabia. *Modelling*, 4(1), 37-55.
- Chou, C. Y., Hsu, D. Y., & Chou, C. H. (2023). Predicting the onset of diabetes with machine learning methods. *Journal of Personalized Medicine*, 13(3), 406.
- Desai, G. S., Hajare, S., & Kharde, S. (2023). Prevalence of non-alcoholic steatohepatitis in a general population of North Karnataka. *The Scientific Temper*, 14(04), 1106-1112.
- Eid, M. M., El-Kenawy, E. S. M., Khodadadi, N., Mirjalili, S., Khodadadi, E., Abotaleb, M., ... & Khafaga, D. S. (2022). Meta-heuristic optimization of LSTM-based deep network for boosting the prediction of monkeypox cases. *Mathematics*, 10(20), 3845.
- El-Kenawy, E. S. M., Abdelhamid, A. A., Ibrahim, A., Mirjalili, S., Khodadadi, N., Alduailij, M. A., ... & Khafaga, D. S. (2023). Al-Biruni Earth Radius (BER) Metaheuristic Search Optimization Algorithm. *Comput. Syst. Sci. Eng.*, 45(2), 1917-1934.
- Ellulu, M. S., & Samouda, H. (2022). Clinical and biological risk factors associated with inflammation in patients with type 2 diabetes mellitus. *BMC Endocrine Disorders*, 22(1), 16.
- Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216, 21-30.
- Forbus, J. J., & Berleant, D. (2022). Discrete-event simulation in healthcare settings: a review. *Modelling*, 3(4), 417-433.
- Katarya, R., & Polipireddy, S. (2020). Identifying Risks in Cardiovascular Disease Using Supervised Machine Learning Algorithms.
- Khafaga, D. (2022). Meta-heuristics for feature selection and classification in diagnostic breast cancer. *Computers, Materials and Continua*, 73(1), 749-765.
- Mahmood, A., Irfan, A., & Wang, J. L. (2022). Developing efficient small molecule acceptors with sp²-hybridized nitrogen at different positions by density functional theory calculations, molecular dynamics simulations and machine learning.

Chemistry–A European Journal, 28(2), e202103712.

Mahmood, A., Irfan, A., & Wang, J. L. (2022). Machine learning for organic photovoltaic polymers: a minireview. *Chinese Journal of Polymer Science*, 40(8), 870-876.

Oubelaid, A., Shams, M. Y., & Abotaleb, M. (2022). Energy efficiency

modeling using whale optimization algorithm and ensemble model. *J. Artif. Intell. Metaheuristics*, 2(1), 27-35.

Shazly, K., & Khodadadi, N. (2023). Credit card clients classification using hybrid guided wheel with particle swarm optimized for voting ensemble. *J. Artif. Intell. Metaheuristics*, 2, 46-54.