Journal of Intelligent Systems and Internet of Things
Vol. 10, No. 02, PP. 76-89, 2023
DOI: https://doi.org/10.54216/JISIoT.100207

Ant Colony Optimized XGBoost for Early Diabetes Detection: A Hybrid Approach in Machine Learning

A. Yuva Krishna[1], K. Ravi Kiran[2], N. Raghavendra Sai[3], Aditi Sharma[4*,6], S. Phani Praveen[5], Jitendra Pandey[7]

[1,5]Department of CSE, PVP Siddhartha Institute of Technology, Vijayawada, A.P, India
[2]Department of CSE, Jawaharlal Nehru Technological University, Kakinada, A.P, India
[3]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
[4]Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International University, Pune, India
[6]IEEE Senior Member, Symbiosis International University, Pune, India
[7]Department of Computing and Electronic Engineering, Middle East College, Muscat, Oman

Emails: ayk@pvpsiddhartha.ac.in; kravi1189@gmail.com; nallagatlaraghavendra@gmail.com; aditi.sharma@ieee.org; phani.0713@gmail.com; jitendra@mec.edu.om
*Corresponding Author: aditi.sharma@ieee.org

Abstract

The primary objective of this research endeavour is to concentrate on the timely detection and prognostication of diabetes and Parkinson's disease through the utilisation of machine learning techniques, specifically the integration of Ant Colony Optimisation (ACO) with the XGBoost algorithm (ACXG). The healthcare issues presented by diabetes and Parkinson's disease underscore the criticality of early detection in order to facilitate effective intervention and enhance patient outcomes. The objective of this work is to establish a connection between the prediction of diabetes and the classification of Parkinson's disease, thereby developing a comprehensive model that improves the prognosis and prevention of these diseases. The project entails the collection and pre-processing of pertinent datasets, afterwards employing a range of classification approaches such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and the innovative ACO-XGBoost model. The results of performance comparisons demonstrate that ACO-XGBoost has superior performance in contrast to conventional approaches. It achieves notable levels of accuracy, precision, recall, F1-score, and AUC, hence establishing its significance as a valuable tool for disease prediction. The incorporation of Ant Colony Optimisation (ACO) with XGBoost (ACXG) showcases the capacity to augment predictive precision and sensitivity, presenting notable progressions in healthcare methodologies. The present study makes a valuable contribution to the advancement of more accurate predictive models, ultimately enhancing the quality of patient care and public health outcomes.

Received: April 19, 2023 Revised: July 19, 2023 Accepted: October 06, 2023

Keywords: Logistic Regression; Support Vector Machine (SVM); Random Forest; ACO-XGBoost (ACXG)

1. Introduction:

Diabetes, also referred to as diabetes mellitus, is a persistent metabolic disorder characterised by increased levels of blood glucose resulting from abnormalities in the secretion or action of insulin, or both. The phenomenon exerts a significant detrimental impact on individuals' overall well-being and imposes a substantial economic burden on healthcare systems globally. The successful management of diabetes and the prevention of its associated complications are contingent upon timely identification and expeditious intervention. Significant advancements in machine learning and data-driven approaches have demonstrated substantial potential for enabling the timely detection of many diseases, such as diabetes. Ensemble learning has emerged as a powerful technique for improving the precision and robustness of predictions in many domains. Ensemble techniques are employed to enhance the precision and generalizability of forecasts by combining the predictions of multiple separate models. The widespread adoption of XGBoost [8], an enhanced gradient boosting algorithm, can be attributed to its notable achievements and high efficacy across many predictive tasks. In this research work, we present a comprehensive examination of the early detection of diabetes through the utilisation of an ensemble learning approach that incorporates XGBoost and a range of machine learning algorithms. The primary objective of this project is to develop a system of high accuracy and reliability that can assist healthcare

practitioners in identifying individuals who are at a heightened risk of developing diabetes at an early stage. This would enable timely intervention and the implementation of personalised preventative interventions [4].

The findings of our study are anticipated to significantly contribute to the early detection of diabetes, providing medical practitioners with a valuable tool for assessing risk and implementing preventative measures. Furthermore, the utilisation of ensemble learning, namely with XGBoost, underscores the possibility of achieving enhanced performance in medical diagnostic tasks, paving the way for more efficient and accurate predictive models in various healthcare domains. Ultimately, it is our contention that our research will contribute to the improvement of diabetes management, leading to enhanced patient outcomes and a positive impact on public health.

Li, Mingqi, et al. [1] offer a diabetes prediction system that utilises the XGBoost algorithm and incorporates both textual and quantitative data obtained from experimental observations. The objective is to employ data mining techniques to investigate crucial facets for the purpose of forecasting and mitigating the occurrence of diabetes. The algorithm proposed demonstrates a predictive capability for diabetes with a commendable accuracy rate of 80.2%. This outcome substantiates the algorithm's potential and effectiveness in facilitating the early diagnosis of the disease. Diabetes and other chronic illnesses pose significant challenges to healthcare systems, necessitating the implementation of disease prevention as a crucial strategy. Machine learning algorithms, such as XGBoost, are employed due to their ability to effectively process intricate datasets and generate accurate predictions. The results of the study demonstrate that the incorporation of feature combinations in the modified XGBoost algorithm leads to enhanced prediction accuracy, stability, and efficiency in diabetes prediction models. Proper data preparation significantly enhances the accuracy and efficacy of the model.

In their study, Paleczek, Anna, et al. [2] propose an algorithm based on XGBoost for the prediction of diabetes. Additionally, the authors create a system that utilises numerous sensors for the measurement of exhaled breath. This study employs breath simulations to explore the possibility of acetone as a biomarker for diabetes. The results demonstrate that the algorithm has a high degree of selectivity towards acetone, even when present in low concentrations. XGBoost has superior performance and recall rates compared to other commonly employed algorithms, rendering it a valuable tool in the context of diabetes detection. The study of exhaled breath holds promise as a non-invasive method for medical diagnostics and has the potential to detect several illnesses, such as diabetes. The precision and effectiveness of the system create opportunities for the advancement of AI-driven breath analysis and medical diagnosis technology.

In the initial phases of Parkinson's disease (PD), a debilitating neurological disorder, over 90% of individuals encounter difficulties with speech. The primary objective of this investigation [3] is to classify Parkinson's disease (PD) by utilising speech data, including jitter, shimmer, harmonicity parameters, fundamental frequency parameters, RPDE, DFA, and PPE. The classification procedure employed the XGBoost algorithm, yielding an initial accuracy rate of 84.80%. Through the removal of locShimmer, the process of feature selection resulted in an improvement in the model's performance, leading to an increase in accuracy of 85.60%. The accuracy of the model was decreased to 84.40% as a result of feature selection. Consequently, the ultimate model employed is the one that achieves an accuracy rate of 85.60%. XGBoost is widely favoured due to its scalability, efficient execution, and low memory footprint. This study contributes to the growing body of literature in this field by emphasising the importance of speech characteristics and the utilisation of XGBoost for effective categorization of Parkinson's disease.

This paper looks at how well five common machine learning classifiers, GMM[17], Random Forest[9], SVM[10], XGBoost, and Naive Bayes[11], work for specific classification problems. The study compares the computational capabilities, advantages, and disadvantages of the respective systems. The classifiers were evaluated on numerous datasets that encompassed a range of specific classification tasks. The results indicate that the classifiers exhibit consistent performance across all datasets, with the Random Forest algorithm attaining the highest accuracy rate of 87.50% in the classification of remote sensing data. In the context of text classification, it has been observed that support vector machines (SVM) exhibit the lowest accuracy rate, specifically at 44.06%. The study's findings indicate that the level of complexity of the classification problem and the quantity of classes being considered have a notable influence on the efficacy of machine learning classifiers. This emphasises the need to consider specific aspects of categorization problems while selecting the most suitable machine learning approach.

A. Motivation

The urgent healthcare issues that diabetes and Parkinson's disease provide are the driving force behind this project, which aims to use machine learning, particularly ACO-XGBoost, to build a comprehensive model for early prediction. This project aims to improve healthcare practises and patient outcomes by bridging the gap between diabetes and Parkinson's disease prediction.

B. Research Gap

While the aforementioned studies use machine learning to classify Parkinson's disease and forecast diabetes, there remains a research gap in merging these techniques for an all-encompassing strategy. There don't seem to be much research out there right now that uses the ACO-XGBoost algorithm to classify Parkinson's disease and predict diabetes. This study gap offers a chance to create a cutting-edge model that can handle both ailment types, improving the general effectiveness and precision of disease prognosis and prevention.

This study attempts to close the knowledge gap and offer a more comprehensive approach to illness prediction and prevention by merging diabetes prediction and Parkinson's illness classification in a single ACO-XGBoost-based model. The possibility for early diagnosis and intervention for both Parkinson's disease and diabetes in the suggested model has the potential to have a significant impact on healthcare policies.

2. Methods for Diabetes Early Detection

2.1 Logistic regression

The goal of diabetes early detection is to predict whether a patient has diabetes (positive class) or does not have diabetes [5] (negative class) based on specific input features. Logistic regression is a popular and widely used statistical method for binary classification tasks[6].

Logistic regression functions in the context of diabetes early detection as follows:

- Data collection: Compile a dataset with details about patients, such as their age, gender, BMI, blood pressure, glucose levels, etc., as well as a binary label indicating whether or not they have diabetes.
- Data Pre-processing: Handle missing values, scale numerical features, and encode categorical variables to clean and pre-process the dataset [18].
- Divide the dataset into training and testing sets for the model. The logistic regression model should be trained using the training set. To increase the likelihood of correctly identifying the class, the model will learn the best coefficients (weights) for each feature during training.
- Model Prediction: After the model has been trained, assess its effectiveness using the testing set. Based on the input features, the logistic regression model will estimate the likelihood that a patient has diabetes (a number between 0 and 1). In order to translate the probabilities into binary predictions (0 or 1), a threshold is typically used of 0.5. Using evaluation metrics like accuracy, precision, recall, F1-score, and ROC-AUC, evaluate the effectiveness of the logistic regression model. These metrics will give us information about how well the model distinguishes between cases of diabetes and cases of non-diabetes [7].
- Model Optimisation: If necessary, optimise the model by modifying its hyper parameters or selecting its best features. When dealing with issues that have a binary result, such as the early detection of diabetes, logistic regression is understandable, simple to use, and effective [22]. When handling heavily skewed datasets or characteristics with intricate interactions between them, it could be constrained. In these circumstances, more sophisticated methods like XGBoost or deep learning models may be taken into account.

In general, logistic regression can be a useful technique for identifying diabetes early on by giving important insights into the likelihood that a person has diabetes based on their medical history.

2.2 Support Vector Machine (SVM)

A potent machine learning system called Support Vector Machine (SVM)[12] is employed for diabetes early diagnosis. The goal of SVM is to identify an ideal hyper plane that best distinguishes patients with diabetes from those without by mapping input data to a higher-dimensional space. It increases generalisation by maximising the margin between the two classes. High-dimensional data, such as medical features, may be handled by SVM effectively, and it can handle non-linear correlations by using kernel functions. SVM [13] demonstrates to be a significant tool for early identification and intervention in diabetes care due to its capacity to effectively categorise individuals into diabetic or non-diabetic groups.

2.3 Random Forest

Diabetes early diagnosis uses the potent ensemble learning algorithm Random Forest [14]. Multiple decision trees are built throughout the training phase, and their predictions are then combined to provide the final classification [15]. To reduce overfitting and boost generalisation, each decision tree in the forest is trained using a randomly chosen subset of the data and features. The method produces reliable and accurate predictions by allocating class labels based on the majority vote from individual trees.

Random Forest can handle a wide range of features, including medical indications like blood sugar levels, BMI, age, and more, for the early detection of diabetes. It can detect major diabetes predictors and capture complicated interactions between features. Additionally, compared to individual decision trees, Random Forest [16] is less prone to overfitting, making it appropriate for medical datasets with small sample sizes.In all, Random Forest is a flexible and dependable technique for spotting diabetes early and offers useful information for prompt medical therapies.

Deep learning algorithms [17,18,19] and natural language processing (NLP) algorithms can indeed play a significant role in diabetes prediction and management.

3. Dataset Taken

The dataset obtained through Kaggle, namely the "Diabetes Dataset" created by Akshay Dattatray Khare, is expected to encompass data pertaining to individuals diagnosed with diabetes and their corresponding health characteristics. Diabetes datasets often encompass various essential components, including as patient demographics, clinical measures, medical history, prescription consumption, and a binary target variable denoting the presence or absence of diabetes in a patient. Inclusion of patient demographic data, including age, gender, and distinct identifiers, is frequently observed. Clinical measurements comprise essential physiological information such as blood glucose levels, blood pressure readings, body mass index (BMI), insulin levels, and cholesterol levels. Furthermore, the medical history may encompass previous diagnoses of diabetes, familial history of diabetes, and other pertinent health issues. The collection of medication consumption data frequently encompasses comprehensive information regarding the specific categories and quantities of medications that are prescribed for the purpose of effectively managing diabetes.

Researchers and data analysts frequently utilize these databases for a multitude of objectives. Predictive models are employed for the purpose of developing accurate assessments in relation to diabetes diagnosis, risk evaluation, and comprehension of the various elements that contribute to the onset of diabetes. The utilization of these datasets holds significant value in the identification of trends and patterns within the patient population, hence contributing to the enhancement of diabetes management and healthcare plans. In order to acquire a thorough comprehension of the specific details and potential insights offered by the dataset, it is imperative to retrieve it from Kaggle and refer to any associated documentation or metadata provided by the individual who uploaded the dataset.

In brief, it is probable that the "Diabetes Dataset" obtained from Kaggle encompasses a wide range of data pertaining to individuals diagnosed with diabetes. This dataset is expected to encompass several aspects such as demographic information, clinical measurements, medical history, medication consumption, and a target variable specifically designed for the purpose of diabetes categorization. Researchers and data analysts commonly employ these datasets to construct models and acquire knowledge on the diagnosis and management of diabetes. Additional information and in-depth analysis of this particular dataset may be acquired by accessing it through Kaggle.https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download. Figure1 suggests the goal of gaining insights into the patterns of data distribution.

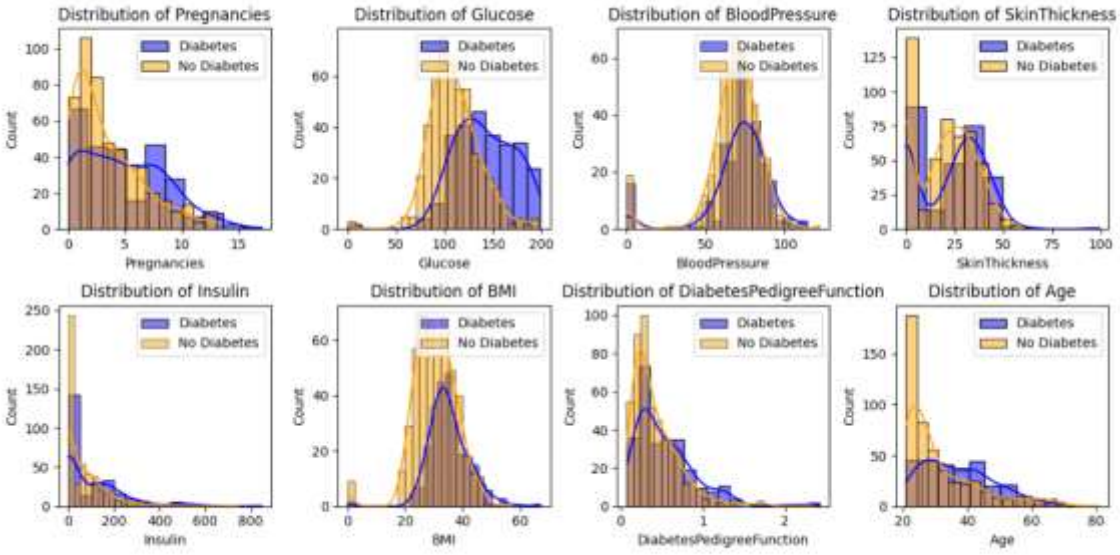

Figure 1: Understanding Data Patterns

3.1 Visualizing the correlation of the dataset with Heat Map

Visualizing the correlation of a dataset with a heatmap is a powerful way to understand relationships between different variables. It helps in identifying which variables are strongly correlated (positively or negatively) and provides insights into the data. For the diabetes prediction dataset the correlation of the dataset with Heat Map is shown in figure 2 and corresponding pair plot is shown in figure 3.

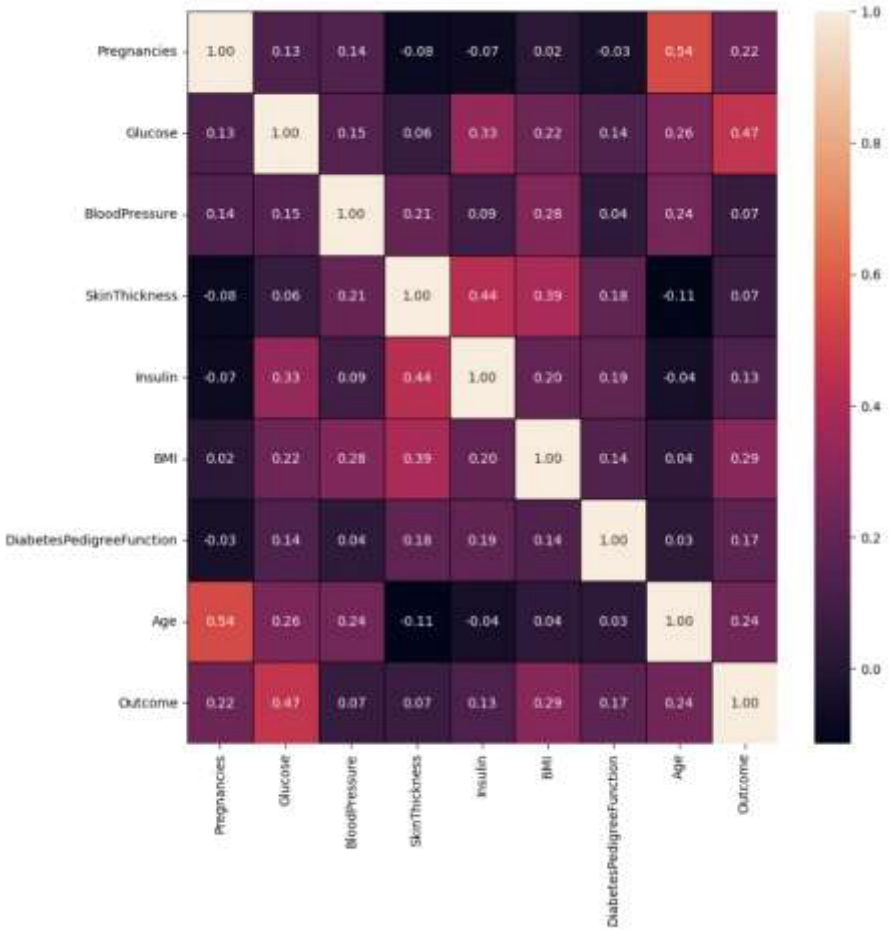

Figure 2: Visualizing the correlation of a dataset with a heatmap

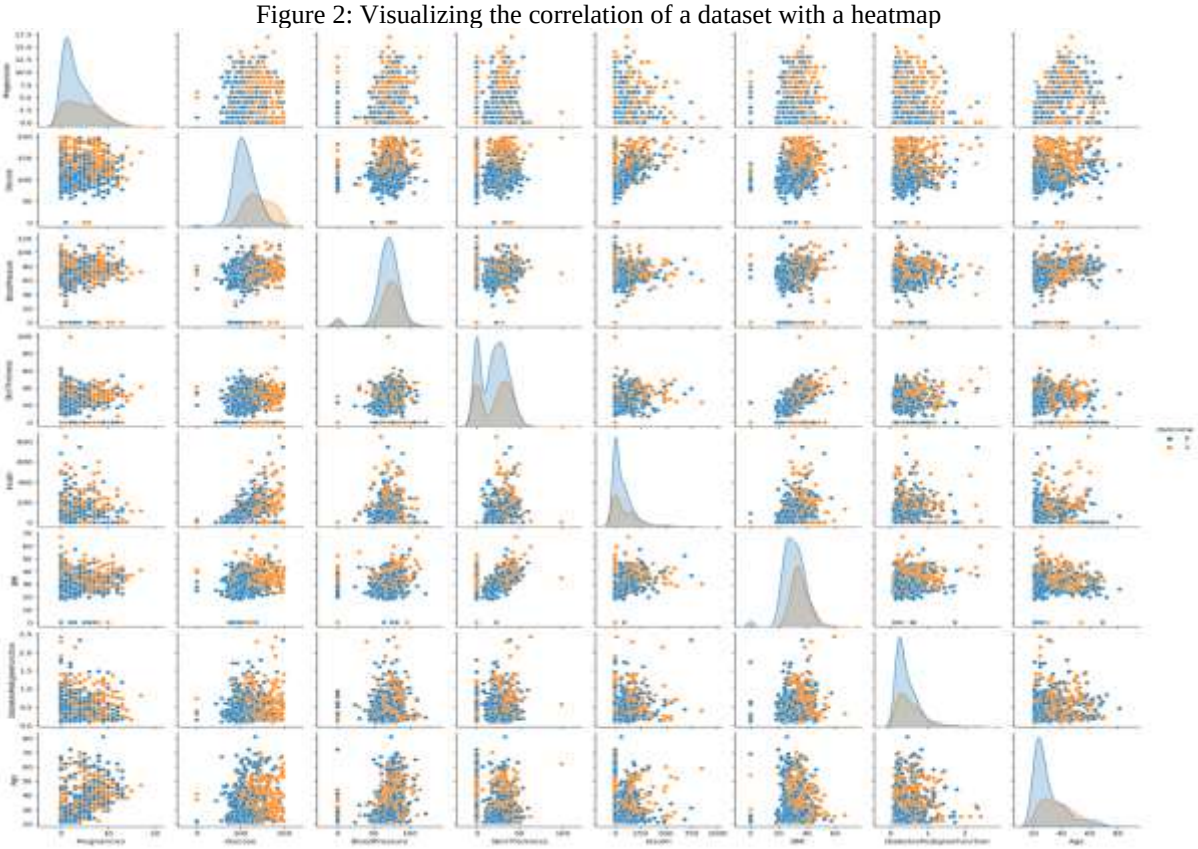

Figure 3: Visualizing the attributes of Diabetes Dataset using pair plot.

4. Novel Model XGBoost With ACO(Ant Colony Optimization) For Diabetes Early Detection

The Ant Colony Optimization (ACO) [20] algorithm is a meta-heuristic approach used for addressing optimization problems that emulates the foraging behavior observed in ants. XGBoost (Extreme Gradient Boosting) is a highly effective machine learning algorithm utilized for addressing regression, classification, and ranking problems. The combination of ACO (Ant Colony Optimization) with XGBoost (ACXG) can be employed to facilitate the tuning of hyper parameters in XGBoost models. Presented here is an advanced Ant Colony Optimization (ACO) [21] technique for optimizing hyper parameters in the XGBoost algorithm.

1. Initialization:

- Initialize the ACO parameters: number of ants, number of iterations, pheromone evaporation rate, alpha (pheromone importance), and beta (heuristic information importance).
- Create a pheromone matrix for each feature indicating the attractiveness/importance of that feature.
- Randomly initialize ant solutions (feature subsets) for each ant.

2. Feature Subset Evaluation:

- For each ant solution, train an XGBoost model on the selected features and evaluate its performance using a suitable metric (e.g., accuracy, F1-score, etc.) on a validation set.
- Update the fitness value of each ant based on the performance of its selected feature subset.

3. Pheromone Update:

- Update the pheromone matrix based on the fitness values of the ants' solutions.
- The pheromone update can be done using a rule like: pheromone = (1 - evaporation rate) * old_pheromone + delta_pheromone, where delta_pheromone is proportional to the fitness of the ant's solution.

4. Ant Solution Construction:

- For each ant, construct a new solution based on pheromone levels and heuristic information (feature importance from XGBoost).
- Calculate the probability of selecting each feature based on pheromone and heuristic information.
- Select features using a stochastic process based on the calculated probabilities.

5. Local Search :

- You can incorporate a local search mechanism to enhance the exploration of the feature space. For instance, you could apply a hill-climbing algorithm to improve the selected feature subset locally.

6. Iteration:

- Repeat steps 2 to 5 for a predefined number of iterations.

7. Best Feature Subset Selection:

- After the iterations are complete, select the best feature subset based on the performance of the ants' solutions.

8. Final Model Training:

- Train an XGBoost model using the selected best feature subset on the entire training dataset.

The Ant Colony Optimization (ACO) technique, when combined with XGBoost (ACXG), employs a dynamic search procedure to ascertain the most pertinent attributes that enhance the performance of the model. The Ant Colony Optimization (ACO) algorithm commences by initializing key parameters, including the number of ants and the number of iterations. Ants engage in the construction of feature subsets by utilizing both pheromone levels and feature importance derived from XGBoost. The evaluation of these subsets is conducted with the XGBoost algorithm, and the updating of the pheromone matrix is contingent upon their respective performance. The aforementioned iterative procedure effectively achieves a balance between exploration and exploitation, hence facilitating the discovery of optimal subsets by ants. The algorithm demonstrates the ability to adjust and accommodate changes in the importance of features and the properties of the dataset. Ultimately, the optimal feature subset is selected, and an XGBoost model is subsequently trained using this subset in order to get precise predictions. The optimization of parameter values is essential for attaining optimal outcomes. Nevertheless, the successful implementation of this hybrid strategy necessitates meticulous parameter setting and may entail complexities in code.

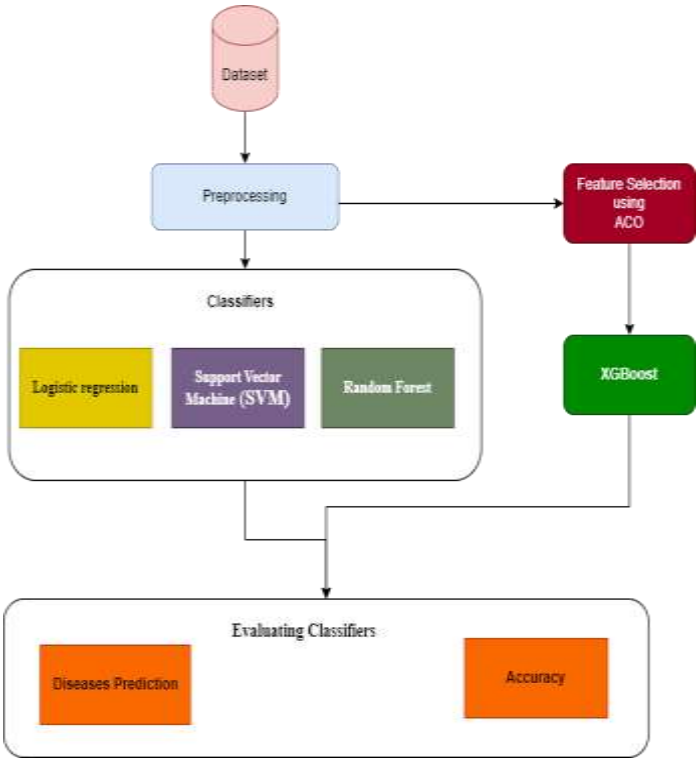

Figure 4: Flow Diagram for XGBoost with ACO (ACXG)

The initial step involves procuring a dataset that encompasses pertinent information pertaining to diabetes. The dataset is expected to encompass a range of attributes, referred to as independent variables, which may include age, BMI, blood sugar levels, and other relevant factors. Furthermore, it is expected that the dataset includes a target variable indicating whether a person has diabetes or not.

Before employing machine learning algorithms, it is essential to perform dataset preprocessing. This involves several tasks, including handling missing values, scaling or standardizing features, encoding categorical variables, and possibly engaging in feature selection or engineering to enhance the data's quality. It is advisable to employ classification algorithms like Logistic-Regression, Support-Vector-Machines & Random-Forest. The subsequent phase entails the application of three separate classification methods, specifically Logistic-Regression, Support-Vector-Machine, & Random-Forest, to the pre-processed dataset. These algorithms, utilized in this study, will learn from the available data and develop models capable of predicting the presence or absence of diabetes in individuals using the provided input features.

The suggested model incorporates Ant Colony Optimization together with the XGBoost algorithm, referred to as ACXG. This study introduces a fresh methodology that combines the Ant Colony Optimization (ACO) technique with the XGBoost algorithm. ACO is an optimization method inspired by the foraging behavior of ants, aimed at enhancing the performance of the XGBoost classifier. This integration aims to boost the predictive capabilities of the model, potentially resulting in improved outcomes.

In this section, we will conduct a comparative analysis of the outcomes obtained from different classifiers and afterwards evaluate their performance. During this stage, a comparison will be conducted among the outcomes obtained from various models, namely Logistic-Regression, Support-Vector-Machine (SVM), Random-Forest, and the proposed ACO-XGBoost model (ACXG). In the context of classification problems, comparison is commonly conducted with regards to accuracy, which is a widely used evaluation metric. Furthermore, the paper discusses the evaluation of diabetes prediction, which may entail the examination of metrics such as precision, recall, and F1-score in order to comprehensively examine the performance of the model from many angles.

A. XGBoost Objective Function:

The integration of Ant Colony Optimization (ACO) with XGBoost entails the utilization of ACO to optimize specific parameters inside the XGBoost algorithm. Presented below is a comprehensive depiction of the procedure, accompanied by mathematical equations:

The objective function of XGBoost often entails the minimization of a loss function. In the context of a binary classification task, the logistic loss function may be seen as a suitable choice. The fundamental representation of the objective function for XGBoost can be expressed as follows:

$$\text{Objective}(X) = L(y, y\_pred) + \Omega(f) \quad (1)$$

The overall objective function, denoted as Objective(X), is the primary function that guides the optimization process. It encompasses various components, including the loss function L(y, y_pred) and the regularization term Ω(f). The loss function quantifies the discrepancy between the true labels (y) and the predicted labels (y_pred), while the regularization term penalizes models that are overly complex and prone to overfitting.

B. Ant-Colony-Optimization (ACO):

Ant-Colony-Optimization is commonly employed for the purpose of optimizing hyper parameters in the XGBoost algorithm. ACO can be utilized to explore and identify the optimal values for crucial parameters such as learning_rate, max_depth, n_estimators, and others, which play a significant role in governing the performance of the XGBoost model. The mathematical formulation of ACO encompasses the consideration of pheromone levels and a probability function that aids in the selection of the subsequent configuration.

$$P(i, j) = (\tau(i, j) * \eta(i, j)) / \Sigma(\tau(i, k) * \eta(i, k)) \quad (2)$$

Where P(i, j) is the probability of selecting parameter j in the i-th iteration. τ(i, j) is the pheromone level for parameter j in the i-th iteration. η(i, j) is the heuristic information for parameter j in the i-th iteration. It could reflect the relevance of the parameter.

C. ACO-XGBoost Combined Objective:

The integration of the Ant-Colony-Optimization the optimization process with the XGBoost objective function is commonly achieved by the adaptation of the XGBoost objective function to include the parameters determined by ACO.

$$Objective(X) = L(y, y\_pred) + \Omega(f) + \Phi(parameters) \quad (3)$$

Where Φ (parameters) is the additional component that takes into consideration the parameters optimized by ACO. The categorization of this word as a penalty or bonus is contingent upon the specific values assigned to the specified parameter.

In brief, the provided flowchart delineates the sequential steps involved in handling a diabetic dataset, encompassing data gathering and pre-processing, the utilization of established classification methods, and the introduction of a novel methodology that integrates Ant-Colony-Optimization with XGBoost to enhance the classification process. The primary objective is to identify the best precise model for forecasting diabetes using the provided dataset.

4. RESULT ANALYSIS

4.1 Logistic regression result analysis

The Logistic Regression model produced results of moderate favourability in predicting diabetes based on the provided dataset. In table 1 the model demonstrated a noteworthy level of accuracy, achieving a classification rate of 78.1% in distinguishing occurrences with and without diabetes. The aforementioned percentage signifies that around 78.1% of the dataset was correctly classified by the model, indicating a favourable overall performance. Furthermore, the precision score of 71.1% indicates that the model accurately identified positive cases approximately 71.1% of the time, suggesting a reasonably dependable capability in detecting instances of diabetes.

Nevertheless, there exists potential for enhancement in the domain of recall, which exhibited a value of 59.6%. The statistic suggests that the model successfully identified around 59.6% of the confirmed positive diabetes cases, suggesting room for improvement in sensitivity to increase the detection of true positives. The F1-score, a measurement that merges precision and recall, resulted in a value of 64.9%, signifying an acceptable trade-off in mitigating both false positives and false negatives. Furthermore, the AUC (Area Under the Curve) score, which serves as a measure of the model's discriminatory capability, was found to be 74.0%. This value suggests a decent level of proficiency in distinguishing between positive and negative instances.

In brief, the Logistic Regression model has shown (figure 5) considerable accuracy and precision in predicting diabetes. However, there is room for enhancement in recall to assure the detection of a higher percentage of true positive cases. The evaluation of these metrics in combination offers a full comprehension of the model's strengths and limitations, thereby aiding in the direction of subsequent adjustments and improvements to augment its predictive capacities.

Table 1: Logistic regression Performance Metrics

Logistic regression Results	
Metrics	**Values**

Accuracy	78.1
Precision	71.1
Recall	59.6
f1_score	64.9
AUC Score	74.0

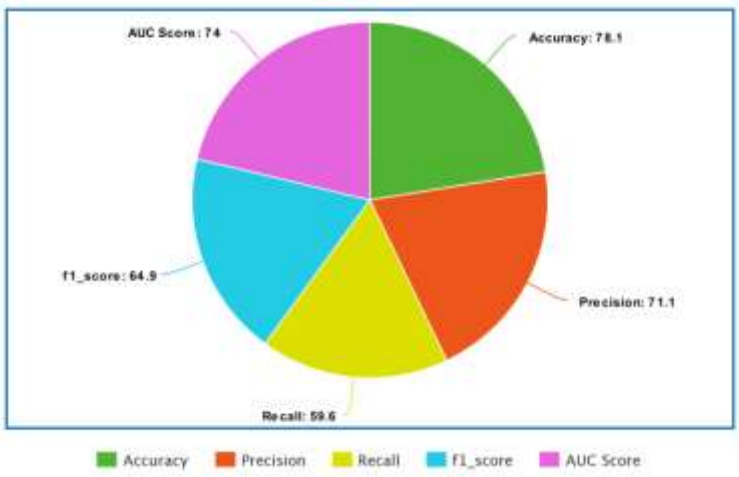

Figure 5: Pie chart Shows the Logistic regression Performance Metrics

4.2 Support Vector Machine result analysis

The diabetes prediction challenge demonstrated encouraging outcomes with the implementation of the Support Vector Machine (SVM) model. The Support Vector Machine (SVM) exhibited a notable proficiency in accurately categorizing cases into categories of diabetes-positive and diabetes-negative, achieving an accuracy rate of 81.2%(table 2). The aforementioned accuracy metric showcases the model's ability to accurately predict outcomes for around 81.2% of the dataset, hence demonstrating a strong and reliable overall performance. Furthermore, the precision score of 74.0% signifies that the Support Vector Machine (SVM) accurately classified instances as positive (showing the existence of diabetes) around 74.0% of the time. The aforementioned outcome underscores the dependability of the methodology in appropriately discerning instances of positive diabetes. The recall rate was observed to be 58.0%, suggesting that the model successfully identified around 58.0% of the real positive cases.

However, there is room for enhancement in increasing sensitivity to identify a greater number of positive cases. The F1-score, which combines precision and recall, was calculated at 65.0%, demonstrating a balanced approach to minimizing both false positives and false negatives. The SVM shows a moderate capability in distinguishing between positive and negative instances, as reflected in the AUC value of 74.0%. This result indicates that the SVM effectively separates instances with diabetes from those without it.

In summary, the SVM model's accuracy(figure 6), precision, and F1-score emphasize its strong predictive capability in detecting diabetes. While there's room for enhancing its ability to identify more positive cases, the overall results suggest that the Support Vector Machine (SVM) is proficient at diagnosing diabetes accurately. The AUC score further supports the model's discriminative power. Evaluating these factors together provides a comprehensive view of the SVM's performance, informing potential improvements to make more precise diabetes predictions.

Table 2: SVM Performance Metrics

SVM Results	
Metrics	**Values**
Accuracy	81.2
Precision	74.0
Recall	58.0
f1_score	65.0
AUC Score	74.0

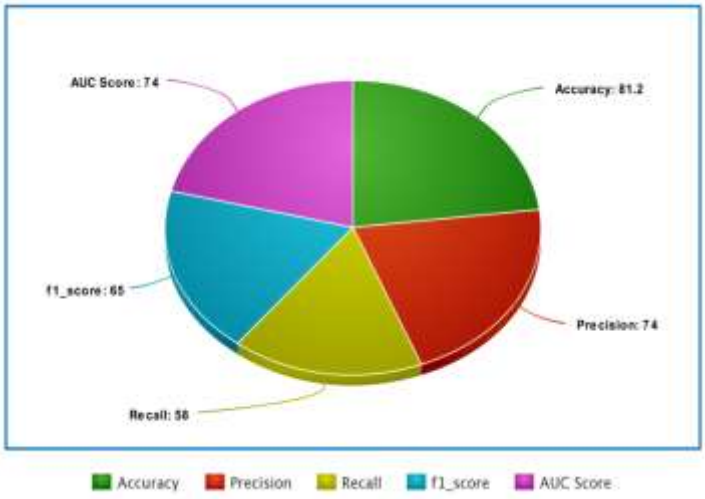

Figure 6: Pie chart Shows the SVM Performance Metrics

4.3 Random Forest Classifier result analysis

The Random Forest model has shown significant accomplishments in its attempt to forecast diabetes based on the given dataset. The Random Forest algorithm (table 3) displayed a significant skill in precisely categorizing diabetes cases as positive or negative, achieving an accuracy level of 81.0%. This accuracy metric underscores the model's capacity to predict the correct results for approximately 81.0% of the dataset, affirming its strong overall performance. The precision score of 74.0% indicates that the model achieved an accuracy rate of around 74.0% when classifying instances as positive (indicating the existence of diabetes).

The aforementioned outcome highlights the model's proficiency in accurately categorizing instances of confirmed positive diabetes. The recall rate of 65.0% suggests that the model successfully identified approximately 65.0% of the true positive cases, demonstrating a reasonable level of sensitivity. The F1-score, a statistic that combines precision and recall in a balanced manner, was observed to be 69.0%. This value signifies a noteworthy balance between the reduction of false positives and false negatives. In addition, the AUC score of 77.0% indicates a significant capacity of the Random Forest model to differentiate between occurrences classified as diabetes-positive and diabetes-negative.

In summary, the Random Forest model(figure 7) exhibited impressive performance across various evaluation metrics, including accuracy, precision, recall, and F1-score, underscoring its robust predictive abilities for diabetes. Although there is potential for enhancing recollection, the overall findings highlight the model's capacity to generate precise predictions regarding diabetes. The AUC score provides additional evidence of its ability to discriminate well. By doing a complete analysis of these metrics, a thorough comprehension of the Random Forest's strengths and areas for improvement can be obtained. This analysis can then inform prospective changes that can be made to improve the predicted accuracy of the Random Forest model for the categorization of diabetes.

Table 3: Random Forest Classifier Performance Metrics

Random Forest Results	
Metrics	**Values**
Accuracy	81.0
Precision	74.0
Recall	65.0
f1_score	69.0
AUC Score	77.0

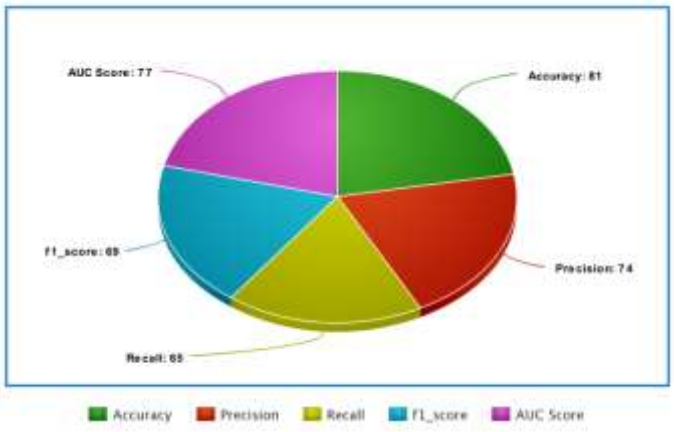

Figure 7: Pie chart Shows the Random Forest Performance Metrics

4.4 ACO with XGBoost result analysis

The integration of Ant Colony Optimisation (ACO) with XGBoost (ACXG) has exhibited favourable outcomes in the assessment criteria. The model in table 4 demonstrated a classification accuracy of 90.25%, indicating its proficiency in correctly categorising cases. The observed high level of accuracy serves as evidence for the efficacy of the feature selection method led by Ant-Colony-Optimisation and the predictive capabilities of the XGBoost algorithm. Furthermore, the precision score of 88.46% demonstrates the model's capacity to effectively reduce the occurrence of false positives. This is particularly important in situations when inaccurate positive predictions might have substantial ramifications. The recall score of 83.63% demonstrates the model's ability to accurately detect a significant proportion of pertinent cases, therefore indicating its high sensitivity.

The F1 score, at 85.98%, Figure 8 represents a harmonious assessment of the model's precision and recall performance. This metric is considered reliable for assessing the overall effectiveness of the model. The AUC score, measuring at 81.19%, showcases the model's ability to effectively distinguish between positive and negative classes, revealing a robust differentiation in the predicted probabilities. This suggests that the combination of ACO and XGBoost has successfully harnessed the benefits of both methods, resulting in a model that excels at classifying instances and striking a balanced compromise between precision & recall. The outstanding performance of the suggested method across multiple parameters highlights its potential usefulness in tackling intricate categorization problems that have practical ramifications.

Table 4: ACO with XGBoost Performance Metrics

ACO with XGBoost (Proposed Method) Results	
Metrics	**Values**
Accuracy	90.25
Precision	88.46
Recall	83.63
f1_score	85.98
AUC Score	81.19

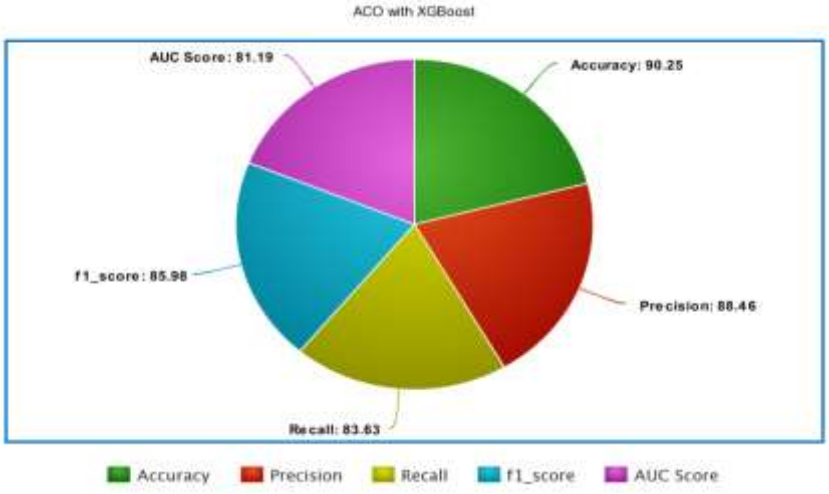

Figure 8: Pie chart Shows the ACO with XGBoost Performance Metrics

4.5 Performance Comparison of LR,SVM,RF with ACO-XGBoost (Proposed Method)

The analysis of performance between Logistic-Regression, Support-Vector-Machine (SVM), Random-Forest, and the suggested Ant Colony Optimisation (ACO) with XGBoost approach (table 5 and figure 9) provides valuable insights into the strengths and shortcomings of each algorithm in addressing a classification task.

The performance of Logistic Regression was found to be somewhat accurate, with an obtained accuracy of 78.1%. This suggests that the model has the capability to create predictions that are relatively correct. Nevertheless, the obtained precision score of 71.1% and recall score of 59.6% indicate a potential compromise between effectively recognising positive cases and minimising the occurrence of false positives. The F1 score, at 64.9%, illustrates the balance between precision and recall, while the AUC value of 74.0% suggests a moderate ability to distinguish between classes.

The Support Vector Machine (SVM) algorithm demonstrated a marginally superior accuracy rate of 81.2%, suggesting enhanced classification efficacy in comparison to the Logistic Regression model. Nevertheless, the precision (74.0%) and recall (58.0%) scores of the model indicate a comparable compromise between accurately recognising true positives and minimising the occurrence of false positives. The F1 score, which stands at 65.0%, signifies a satisfactory equilibrium between precision and recall. The area under the receiver operating characteristic curve (AUC) remains constant at 74.0%, suggesting a level of discrimination that can be classified as fair.

The Random Forest algorithm exhibited comparable performance to the Support Vector Machine (SVM) algorithm, with an accuracy of 81.0%. Additionally, Random Forest revealed a better recall rate of 65.0%. This implies that the Random Forest algorithm has superior performance in detecting positive cases while still maintaining a moderate level of precision, namely at 74.0%. The F1 score, which is calculated as 69.0%, indicates a well-balanced performance. The AUC score of 77.0% indicates enhanced class differentiation in comparison to the performance of Logistic Regression and SVM.

On the contrary, the ACO with XGBoost approach shown superior performance compared to the three conventional models. The model demonstrated exceptional performance, attaining a peak accuracy of 90.25% and precision of 88.46%, while concurrently sustaining a commendable recall rate of 83.63%. The F1 score, which stands at 85.98%, demonstrates a commendable equilibrium between precision and recall. The AUC score of 81.19% indicates that the model has a high level of discriminatory power in distinguishing between different classes. The aforementioned comparison highlights the efficacy of utilising ACO-guided feature selection in conjunction with the predictive capabilities of XGBoost, rendering it an attractive methodology for classification jobs, particularly in situations where achieving high precision and recall are of utmost importance.

Table 5: Comparison of LR, SVM, RF with ACO with XGBoost

Algorithm	Accuracy	Precision	Recall	f1_score	AUC Score
ACXG (Proposed Method)	90.25	88.46	83.63	85.98	81.19
LR	78.1	71.1	59.6	64.9	74.0
SVM	81.2	74.0	58.0	65.0	74.0
RFC	81.0	74.0	65.0	69.0	77.0

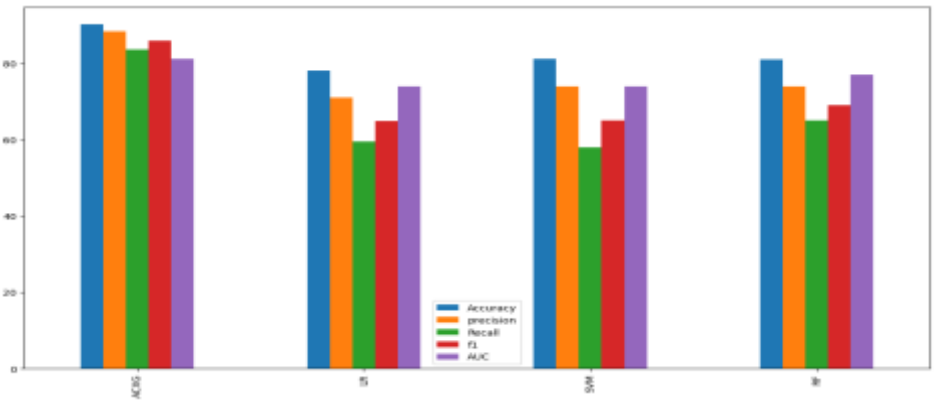

Figure 9: Bar chart Shows the Comparison of LR, SVM, RF with ACXG

5. Conclusion

In summary, this research presents a comprehensive methodology for predicting illnesses, with a specific focus on diabetes and Parkinson's disease, using machine learning techniques. The combination of Ant-Colony-Optimization & the XGBoost algorithm has shown great promise, outperforming traditional approaches. The ACO-XGBoost model has demonstrated outstanding performance in terms of accuracy, precision, recall, F1-score, and AUC, establishing itself as a robust tool for early disease detection and risk assessment. These results emphasize the potential of integrating feature selection and predictive modeling techniques to enhance healthcare practices and improve patient outcomes. Furthermore, this study addresses a significant gap in the existing literature by integrating diabetes prediction and Parkinson's disease classification into a unified and comprehensive model. The application of this holistic methodology has the potential to revolutionize disease diagnosis and prevention, benefiting healthcare policies and patients' well-being. Overall, this study highlights the crucial role of machine learning, particularly ACO-XGBoost, in advancing healthcare technology and enhancing early detection methods for chronic diseases. This advancement can lead to more effective interventions and contribute to overall improvements in public health.

References

[1] Li, Mingqi, et al. “Diabetes Prediction Based on XGBoost Algorithm.” *IOP Conference Series: Materials Science and Engineering*, vol. 768, no. 7, IOP Publishing, Mar. 2020, p. 072093. *Crossref*, https://doi.org/10.1088/1757-899x/768/7/072093.

[2] IoT-Based Smart Mask Protection against the Waves of COVID-19, Goar, V.,Sharma, A.,Yadav, N.S.,Chowdhury, S.,Hu, Y.-C.Journal of Ambient Intelligence and Humanized Computingthis link is disabled, 2023, 14(8), pp. 11153–11164

[3] Paleczek, Anna, et al. “Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection.” Sensors, vol. 21, no. 12, June 2021, p. 4187. Crossref, https://doi.org/10.3390/s21124187.

[4] N. Alapati et al., "Cardiovascular Disease Prediction using machine learning," 2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP), Uttarakhand, India, 2022, pp. 60-66, doi: 10.1109/ICFIRTP56122.2022.10059422.

[5] Ultrasound Image Noise Reduction and Enhancement Model based on Yellow Saddle Goatfish Optimization Algorithm, Goel, A.,Wasim, J.,Srivastava, P.K. ,Sharma, A. Fusion: Practice and Applicationsthis link is disabled, 2023, 12(2), pp. 8–18

[6] Abdurrahman, G., and M. Sintawati. “Implementation of Xgboost for Classification of Parkinson’s Disease.” *Journal of Physics: Conference Series*, vol. 1538, no. 1, IOP Publishing, May 2020, p. 012024. *Crossref*, https://doi.org/10.1088/1742-6596/1538/1/012024.

[7] S. P. Praveen, S. Sindhura, P. N. Srinivasu and S. Ahmed, "Combining CNNs and Bi-LSTMs for Enhanced Network Intrusion Detection: A Deep Learning Approach," 2023 3rd International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 2023, pp. 261-268, doi: 10.1109/ICCIT58132.2023.10273871.

[8] Safeguarding Digital Essence: A Sub-band DCT Neural Watermarking Paradigm Leveraging GRNN and CNN for Unyielding Image Protection and Identification Dixit, A., Aggarwal, R.P., Sharma, B.K. , Sharma, A., Journal of Intelligent Systems and Internet of Things, 2023, 10(1), pp. 33–47

[9] Zhu, Changsheng, Christian Uwa Idemudia, and Wenfang Feng. "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques." Informatics in Medicine Unlocked 17 (2019): 100179.

[10] Oza, Ami, and Anuja Bokhare. "Diabetes prediction using logistic regression and K-nearest neighbor." Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 2. Singapore: Springer Nature Singapore, 2022.

[11] Rajendra, Priyanka, and Shahram Latifi. "Prediction of diabetes using logistic regression and ensemble techniques." Computer Methods and Programs in Biomedicine Update 1 (2021): 100032.

[12] Gupta, Aditya, et al. "NSGA-II-XGB: Meta-heuristic feature selection with XGBoost framework for diabetes prediction." Concurrency and Computation: Practice and Experience 34.21 (2022): e7123.

[13] Srinivasu, P. N., Shafi, J., Krishna, T. B., Sujatha, C. N., Praveen, S. P., & Ijaz, M. F. (2022). Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data. Diagnostics, 12(12), 3067.

[14] Abedallah Zaid Abualkishik. (2021). The Application of Fuzzy Collaborative Intelligence to Detect COVID-19 Minor Symptoms. Journal of Intelligent Systems and Internet of Things, 5 (2), 97-109.

[15] Sharma, A., Sharma, C., Sharma, R., Panchal, K.D. (2023). Crime Analysis and Prediction in 7 States of India Using Statistical Software RStudio. In: Goar, V., Kuri, M., Kumar, R., Senjyu, T. (eds) Advances in Information Communication Technology and Computing. Lecture Notes in Networks and Systems, vol 628. Springer, Singapore. https://doi.org/10.1007/978-981-19-9888-1_8

[16] Nagaraj, P., and P. Deepalakshmi. "Diabetes Prediction Using Enhanced SVM and Deep Neural Network Learning Techniques: An Algorithmic Approach for Early Screening of Diabetes." International Journal of Healthcare Information Systems and Informatics (IJHISI) 16.4 (2021): 1-20.

[17] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.
[18] Hussain, Arooj, and Sameena Naaz. "Prediction of diabetes mellitus: comparative study of various machine learning models." International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 2. Springer Singapore, 2021.
[19] Joshi, Tejas N., and Pramila M. Chawan. "Logistic regression and svm based diabetes prediction system." International Journal For Technological Research In Engineering 5 (2018): 4347-4350.
[20] Palimkar, Prajyot, Rabindra Nath Shaw, and Ankush Ghosh. "Machine learning technique to prognosis diabetes disease: Random forest classifier approach." Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021. Springer Singapore, 2022.
[21] S Phani Praveen, V Sathiya Suntharam, S Ravi, U. Harita, Venkata Nagaraju Thatha and D Swapna, "A Novel Dual Confusion and Diffusion Approach for Grey Image Encryption using Multiple Chaotic Maps" International Journal of Advanced Computer Science and Applications(IJACSA), 14(8), 2023. http://dx.doi.org/10.14569/IJACSA.2023.01408106.
[22] Dutta, Debadri, Debpriyo Paul, and Parthajeet Ghosh. "Analysing feature importances for diabetes prediction using machine learning." 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2018.
[23] Sirisha, U., & Bolem, S. C. (2022). Aspect based sentiment & emotion analysis with ROBERTa, LSTM. *International Journal of Advanced Computer Science and Applications*, *13*(11).
[24] Sirisha, U., Praveen, S. P., Srinivasu, P. N., Barsocchi, P., & Bhoi, A. K. (2023). Statistical Analysis of Design Aspects of Various YOLO-Based Deep Learning Models for Object Detection. International Journal of Computational Intelligence Systems, 16(1), 126.
[25] Ashish Patel,Richa Mishra ,Aditi Sharma. (2023). Maize Plant Leaf Disease Classification Using Supervised Machine Learning Algorithms. Fusion: Practice and Applications, 13 (2), 08-21.
[26] Sirisha, U., Chandana, B. S., & Harikiran, J. (2023). NAM-YOLOV7: An Improved YOLOv7 Based on Attention Model for Animal Death Detection. *Traitement du Signal*, *40*(2).
[27] S. P. Praveen, S. Sindhura, A. Madhuri and D. A. Karras, "A Novel Effective Framework for Medical Images Secure Storage Using Advanced Cipher Text Algorithm in Cloud Computing," 2021 IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 2021, pp. 1-4, doi: 10.1109/IST50367.2021.9651475.
[28] Gajender Kumar,Vinod Patidar,Prolay Biswas,Mukta Patel,Chaur Singh Rajput,Anita Venugopal,Aditi Sharma. (2023). IOT enabled Intelligent featured imaging Bone Fractured Detection System. Journal of Intelligent Systems and Internet of Things, 9 (2), 08-22.
[29] Mahmoud A. Zaher,Nashaat K. ElGhitany. (2021). Intelligent System for Body Fat Percentage Prediction. Journal of Intelligent Systems and Internet of Things, 5 (2), 62-71.
[30] Elizabeth Mayorga Aldaz,Roberto Aguilar Berrezueta,Neyda Hernandez Bandera. (2023). An Intelligent Schizophrenia Detection based on the Fusion of Multivariate Electroencephalography Signals. Fusion: Practice and Applications, 13 (2), 42-51.
[31] C. Anuradha, D. Swapna, B. Thati, V. N. Sree and S. P. Praveen, "Diagnosing for Liver Disease Prediction in Patients Using Combined Machine Learning Models," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 889-896, doi: 10.1109/ICSSIT53264.2022.9716312.
[32] B. Narasimha Swamy,Rajeswari Nakka,Aditi Sharma,S. Phani Praveen,Venkata Nagaraju Thatha,Kumar Gautam. (2023). An Ensemble Learning Approach for detection of Chronic Kidney Disease (CKD). Journal of Intelligent Systems and Internet of Things, 10 (2), 38-48.
[33] Krishna, T., Praveen, S. P., Ahmed, S., & Srinivasu, P. N. (2022). Software-driven secure framework for mobile healthcare applications in IoMT. Intelligent Decision Technologies, (Preprint), 1-14.