Pakistan Journal of Statistics and Operation Research

Hybrid Approach Based on the CHAID Algorithm for Improving Classification Performance of Diabetes Data

Walaa Mohamed Elaraby Mohamed Shallan[1*]

* Corresponding Author

1. Department of Statistics, Mathematics, and Insurance, Faculty of Business, Ain Shams University, Cairo, Egypt, *walaaelaraby@bus.asu.edu.eg*

Abstract

Diabetes, a chronic disease that is becoming more prevalent, presents increasing challenges, especially in low- and middle-income countries, where it is a growing burden. Egypt is the 9th most prevalent country for diabetes in the world, with estimated diabetes prevalence among adults at 15.2%, which raises urgent implications for early detection to limit complications including retinopathy, renal impairment and limb amputation. This study proposes a method to address classification of Type 2 diabetes (T2DM) through implementing and exploring the application of five machine learning algorithms: support vector machine (SVM), naïve Bayes (NB), K-Nearest Neighbor (KNN), Bayesian network (BNC) and stochastic gradient descent (SGD), along with CHAID algorithm to produce conditional segmentation variable to model non-linear interactions while improving expressivity of features used. CHAID analyses found that the best predictor of T2DM involved high levels of the hemoglobin A1c, and insulin resistance. The next best predictors were triglycerides and then followed by age, obesity, and blood pressure. Effects from the metabolic, cardiovascular, and lifestyle variables were small-to-moderate showing a significant amount of clustering. The hybrid model was developed as protection against overfitting, thus allowing robust and generalizable classification performance. The proposed hybrid models outperformed that of a single model. Specifically, SVM via CHAID and SGD via CHAID were able to obtain a perfect classification accuracy of 100% revealing the model's potential as a powerful tool for early detection and examination of risk of diabetes.

Keywords: CHAID Algorithm, Diabetes, Support Vector Machine, Naïve Bayes, K-Nearest Neighbor, Bayesian Network, Stochastic Gradient Descent.

Mathematical Subject Classification: 62H30; 62P10

1. Introduction

Chronic high blood sugar is indicative of diabetes, a long-lasting disease. It may happen due to either a lack of insulin production or the body's inability to use insulin properly. If a person is not diagnosed and treated in a timely manner, the condition can progressively worsen without any symptoms and damage important organs. Devastating outcomes may result, such as diabetic retinopathy-induced blindness, neuropathy, kidney failure, stroke, and other cardiovascular issues (Ong et al., 2023). This trend is also reflected in a parallel rise in prediabetes, which is defined as elevated blood-glucose levels without meeting diabetes criteria and has a proclivity toward the development of T2DM. There is also an alarming global incidence of prediabetes and estimates place it at 992 million cases in 2030 and an astounding 1.171 billion in 2045. This establishes a gap - and increasing- demand for effective early detection and intervention.

The health crisis is staggering. That means in 2021, an estimated 537 million adults lived with diabetes globally, and 90% of them have type 2 diabetes (T2DM). The disturbing part is that half of them, at 44.7% of people with diabetes, were undiagnosed (Ong et al., 2023). Projecting the increase in people with diabetes, we expect to see 643 million with diabetes in 2030, and 783 million in 2045

Most people living with diabetes live in low-above-middle-income countries. The International Diabetes Federation (IDF) estimates the prevalence is 15.2% of adults in Egypt to be diabetic, positioning it the 9th globally, with projections that hypothetically place Egypt in a similar predicament. The economic impact of T2DM in Egypt was USD 1.29 billion in 2010 and is expected to double by 2030(Hegazi, El-Gamal, Abdel-Hady, & Hamdy, 2015).

Egypt provides primary, preventive, and curative healthcare, but there is no formal referral system for patients, and diabetes costs are the lowest in the Middle East and North Africa (MENA) at USD 160–3,000 annually per patient, compared with USD 2,000–7,000 in developed countries. Early detection of Type 2 diabetes mellitus (T2DM) has become necessary due to the high undetected prevalence of diabetes, rising prevalence, and early complication risk of microvascular events before diagnosis. Early detection has been highlighted by the World Health Organization (WHO) and International Diabetes Federation (IDF) since 2003 (Azami & Ebrahimy, 2023; Farag et al., 2023).

As T2DM prevalence increases, the need for accurate statistical modeling for early detection and prevention initiatives becomes vital. Classification algorithms are frequently used in medicine to categorize data according to defined parameters, and many machine learning methods have been advantageous for predicting and categorizing chronic disease like diabetes mellitus (DM) (Abnoosian, Farnoosh, & Behzadi, 2023).

Combining different machine learning models can yield better performance than single models by reducing individual limitations and leveraging complementary mechanisms. In hybrid intelligent systems, each layer contributes additional information, and the system's overall performance depends on the correct operation of all layers. Given that no single classification technique consistently produces the best results, hybrid approaches have been proposed to improve prognosis accuracy, identify those most at risk based on physiological and genetic characteristics, and enhance classification accuracy (Yoon, 2012). Such approaches are increasingly used to support patient monitoring and early disease detection, with Studyers developing various hybrid models and algorithms to overcome the limitations of single models and improve decision making (X. Li, Zhang, & Safara, 2023)

This paper progresses by analyzing recent studies in this field, highlighting the most efficient algorithm with the best performance.

(Murthy & Arunadevi, 2021) sought to enhance an automated procedure for detecting retinal blood vessels (RBVs) relative to diabetic retinal disease (DR), both the segmentation and classification of the impacted vessels. They used Kinetic Gas Molecule Optimization using a centroid to initialize fuzzy means clustering and following that was the classification with a convolutional neural network with bidirectional long short-term memory (CNN with Bi-LSTM). A self-attention mechanism helped to optimize the Bi-LSTM learning, showed better accuracy, specificity, and sensitivity compared to other methods, achieving a mean classification accuracy of 97.09%. (Gashti & Sciences, 2017) sought to enhance support vector machine (SVM) performance and did so by incorporating it with a decision tree (Iterative Dichotomiser 3). The hybrid provided a domain classification accuracy of 0.9125 and performed better than either model alone. In the same vein (Bharadwaj & Minz, 2012) incorporated support vector machines within a C4.5 decision tree as pre-pruning for a more accurate and efficient hybrid classifier. (Haq, Li, Memon, Nazir, & Sun, 2018), created a hybrid system for heart disease diagnosis using logistic regression, K-nearest neighbors (KNN), artificial neural networks (ANN), SVM, Naive Bayes (NB), decision trees, and random forests, as well as feature selection methods like Relief, minimum redundancy maximum relevance, and least absolute shrinkage. The best performance was achieved with Relief with logistic regression (89% accuracy). (Lai, Huang, Keshavjee, Guergachi, & Gao, 2019) developed a diabetes classification model using local Gaussian regression (LGR) and gradient boosting machines. They found LGR and gradient boosting machines achieved a higher sensitivity compared to random forests (RF) and decision trees (DT). (Çalişir & Doğantekin, 2011) developed an SVM model combining linear discriminant analysis as an estimator on a Morlet wavelet kernel model. They reported a successful classification accuracy of 89.74 percent for diabetes diagnostic purposes. (Hassan Yaseen & Alibraheemi, 2022) proposed a COVID-19 detection method based on chest X-rays and CT scans through the use of CNN, KNN, RF, and SVM. The CNN-SVM hybrid achieved the highest accuracy at 99.51% for the X-ray dataset.(Sneha & Gangil, 2019) proposed the use of the Farthest First (FF) clustering algorithm in combination with a Sequential Minimal Optimization (SMO) classifier to predict diabetes using the Pima Indians dataset, achieving 99.4% accuracy.(G. Li, Chen, Yang, & Protection, 2022) proposed a combination of variational mode decomposition, extreme learning machines, and ARIMA for estimating cumulative COVID-19 cases called the GVMD-ELM-ARIMA hybrid model. Compared to other methods, GVMD-ELM-ARIMA demonstrated the highest forecasting accuracy. (Sheela & Arun, 2022) proposed a hybrid PSO-SVM method for detecting new COVID-19 by using MRI images and attained up to 95.78% accuracy above established methods. (Mowafy, Shallan, & Science, 2021) demonstrated how complexity, such as combining multiple correspondence analysis (MCA) and principal component analysis (PCA) with fuzzy c-means (FCM) clustering, producing the highest classification analysis via the MLP via FCM-MCA. (Avula & Asha, 2018) combined the JRIP and Naive Bayes algorithms using a number of medical datasets to improve prediction accuracy. (Mutlu & Acı, 2022) proposed a

parallel-hybrid SVM-SMO-SGD method which again outperformed classical SMO-based methods over dataset such as diabetes stroke prediction, adult health data.

Prior literature reviews have demonstrated that hybrid machine-learning methods are frequently better at predicting than standalone models; however, there is limited application of hybrid machine-learning models and methods on Egyptian diabetes data. This study addresses that limitation by employing Chi-square Automatic Interaction Detection (CHAID)-based segmentation to develop multiple classifiers (SVM, Naïve Bayes, KNN, BNC, SGD) from the data in both private and public hospitals. The study combines CHAID's unique modelling of interactions among variables and the assumption of standard classifiers, allowing for pre- and post-CHAID model comparison using the eight-performance metrics. In its ability to provide one of the first locally contextualized, data-driven T2DM risk hierarchies in Egypt, the study contributes methodological innovation and a substantive clinical application that could be integrated directly into electronic health records. The structure of the paper is as follows: section 2 provides a review of hybrid model applications; section 3 details methods and materials including data collection and preparation methods; section 4 describes model development and evaluation; Section 5 discusses implications, and Section 6 outlines limitations of the Study, future Study directions, conclusions and recommendations.

2. Machine Learning Classifiers

2.1CHAID algorithm

One of the most common decision tree techniques is the Chi-square Automatic Interaction Detection (CHAID) algorithm, which generates a tree structure by repeatedly splitting subsets of the data space into two or more sub-nodes, starting from the complete dataset. At each node, the algorithm determines the best split by combining any available pair of categories of the independent variable until no statistically significant difference remains between the pair with respect to the dependent variable. Interactions between independent variables, which can be directly observed from the tree structure, are naturally handled by the CHAID algorithm. The final nodes define subgroups characterized by different sets of independent variables.

Cross-validation in a CHAID model is through splitting the sample into smaller subsamples. With each of the subsamples being withheld in turn from the development of the tree, the tree is then tested against the omitted subsample as an estimate of misclassification risk. The average risk across all trees provides the cross-validated risk estimate for the entire model (Türe, Kurt, & Kürüm, 2007), The tree construction stops when no significant chi-square value is found between the target variable and the predictors. Consequently, nodes with higher chi-square values appear closer to the root, while terminal nodes have the lowest chi-square values (Wang, Ni, & Stone, 2018).

2.2Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a powerful tool for supervised machine learning algorithms that can be used for both regression and classification and are suited for both linear and non-linear problems. The main goal of SVM is to find a hyperplane in a multidimensional space that divides the feature space into two distinct groups. In SVM, a new subject is subsequently categorized according to which side of the hyperplane is on. SVM separates data by labels. A kernel trick is used to match new data to the best of previously trained data to predict an unknown target label. Hyperplanes are used to separate binary classification cases(Gashti & Sciences, 2017).

$$\boldsymbol{w^T \alpha + b = 0} \tag{1}$$

where denotes a dimensional coefficient vector and denotes an offset value from the original dataset value.

The solution in the linear case is obtained by introducing Lagrangian multipliers, and support vectors are used as data points at the boundaries (Doppala, Bhattacharyya, Chakkravarthy, Kim, & Databases, 2021).

$$\boldsymbol{w = \sum_{i}^{n} \propto_i Y_i X_i} \tag{2}$$

Where n is the number of vectors, Y_i Set the target labels to x,The linear discriminant function is denoted as follows:

$$\mathbf{g(x) = sgn \sum_{i}^{n} \left(\propto_i Y_i X_i^T + b \right)} \tag{3}$$

The decision function for the kernel trick is:

$$g(x) = sgn\sum_{i}^{n}(\propto_i Y_i K(X_i + X) + b) \quad (4)$$

2.3Naïve Bayes Classifier (NB)

The Naïve Bayesian Classifier approach is based on the Bayesian theorem and is especially useful when the inputs have high dimensional. Based on the input, the Bayesian classifier can calculate the most possible output. It is also feasible to input new raw data and improve the probabilistic classifier during runtime. The presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature when using a naive Bayes classification. Bayes' theorem provides a way of calculating the posterior probability $P(c\ /\ x)$, from $P(c), P(x)$,and $P(x|c)$. Naive Bayes classifier considers that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

$$P(c\ /\ x) = P(x\ /\ c)P(c)/P(x) \quad (5)$$

Where $P(c\ /\ x)$ is posterior probability of class (target) given predictor (attribute) of class, $P(x\ /\ c)$ is the likelihood, which is the probability of predictor of given class, $P(c)$ is the class prior probability, $P(x)$ is the predictor prior probability of class(Nikam & Technology, 2015).

2.4K-Nearest Neighbor (KNN)

KNN (K-Nearest Neighbors) is an example of a supervised machine learning algorithm and can be used for either classification or regression tasks. KNN is also considered a "lazy learner" because while it records all the training data, it builds no explicit model during training, meaning it merely saves the entire dataset, and classifies a new point by finding previous instances with similar features. Classification is based on similarity, and the new point is assigned to the category of the closest, stored example (Saxena, 2021).

2.5Bayesian network classifier (BNC)

Bayesian classifiers (BC) are statistical classifiers that predict class membership using probabilities, such as the probability that a given sample belongs to a specific class. Because of their simplicity, computational efficiency, and particularly good performance, Bayesian networks are more commonly used in real-world applications.
The idea behind a Bayesian classifier is that the role of a (natural) class is to predict the values of features for members of that class. Bayesian classifiers are based on the Bayes theorem, which states that:

$$P(c_j/d) = p(d/c_j)\,p(c_j)p(d) \quad (6)$$

$P(c_j/d)$= probability of instance d being in class c_j ,
$p(d/c_j)$= probability of generating instance d given class c_j ,
$p(c_j)$= probability of occurrence of class c_j ,
$p\ (d)$ = probability of instance d occurring (Anuradha & Velmurugan, 2015).

2.6Stochastic Gradient Descent (SGD)

The stochastic gradient descent (SGD) model is a very efficient, simple, and appropriate method for linear classifiers with convex loss functions. The classifier uses regularized linear models in the SGD, the loss gradient is appreciated for each sample at the same time, and the module is updated using the learning rate{Atiyah, 2022 #156}. SGD is an iterative strategy for improving a target function that has reasonable perfection properties (such as being differentiable or sub-differentiable). The factual assessment considers the issue of limiting a target work having the type of an aggregate as shown in Eq. (9), where the boundary (W) that limits $Q(w)$ is to be assessed. Every summand work (Q_i) connected with the $(I - th)$ perception(Grwany, 2021).

$$Q(W) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w) \quad (7)$$

3. Proposed method

The proposed hybrid CHAID classification methodology has a specific workflow which is summarized in Table 1 and illustrated in the flow diagram of Figure 1. Table 1 provides a summary of the individual activities, from data preprocessing to data evaluation, and identifies the tools that could be employed at each phase. Figure 1 exploits the summary in Table 1 and describes the implementation flow.

Table 1 Proposed Hybrid CHAID–ML Classification Methodology

Step	Description	Tool/Software
1. Data Preprocessing	Import and prepare the T2DM dataset.	SPSS V29
2. CHAID Segmentation	Apply the CHAID algorithm to generate a segmentation model and assign each observation to a terminal node group.	SPSS V29
3. Feature Construction	Construct the new categorical feature CHAID_Segment capturing the complex interaction patterns identified by CHAID.	Output of CHAID
4. Classifier Training	Train the machine learning classifiers (SVM, NB, KNN, BNC, SGD) using the CHAID_Segment as the target variable.	WEKA V3.8.6
5. Cross-Validation	Employ a nested 10-fold cross-validation protocol to ensure robustness and prevent data leakage.	WEKA / SPSS
6. Performance Evaluation	Assess and compare the classification accuracy (such as accuracy, precision, recall, F1-score ROC-AUC) of the hybrid models to the baseline models measuring the CHAID assigned labels as well as the original clinical diagnoses.	Results presented in Table 5 (Confusion Matrices), Table 6 (Performance Metrics), and Figures 2, 3, 4 (Visual Comparisons)

3.1 CHAID Segmentation and Hybrid Feature Construction

"In order to provide a clear statistical rationale and technical understanding of the CHAID-derived feature included in this pipeline, we describe its construction and integration below(Zheng & Casari, 2018)

The hybrid methodology employed in this pipeline uses the CHAID algorithm, not as a separate classifier, but rather as an engineering and segmentation process tool, to improve predictive accuracy.

CHAID Segmentation

The original dataset (with target variable $Y_{original}$ {0 = Prediabetics, 1 = Type 2 Diabetics}) is segmented using CHAID based on statistically significant relationships between predictors and the target. The process continues until the terminal nodes are homogenous, making it possible to identify unique subgroups with different risk scenarios(Kass, 1980).

CHAID-derived Segment

A new categorical feature, $CHAID\_Segment$, is created by assigning each observation with a label corresponding to its terminal node. The labels (1, 2) are a label for each statistical segment and do not recode the original target. The label '1' may be made up predominantly of Prediabetics, while '2' is made up predominantly of Type 2 Diabetics. This variable is intended to capture potentially complicated, higher order interactions among predictors(Zheng & Casari, 2018).

Hybrid Feature Space and Classification & Rationale and Performance

$CHAID\_Segment$ was added to the original features:

$[X_{original}, CHAID\_Segment]$

This associated data, with all measurements and interactions derived from CHAID, will be used to train the classifiers (SVM, KNN, Naïve Bayes, etc...).

This gives classifiers the benefit of using patterns already identified, reduces overfitting, and improves interpretability, robustness and accuracy. The CHAID-derived segment provided a new perspective on classifying Type 2 Diabetes while staying within our process(Rokach, 2010).

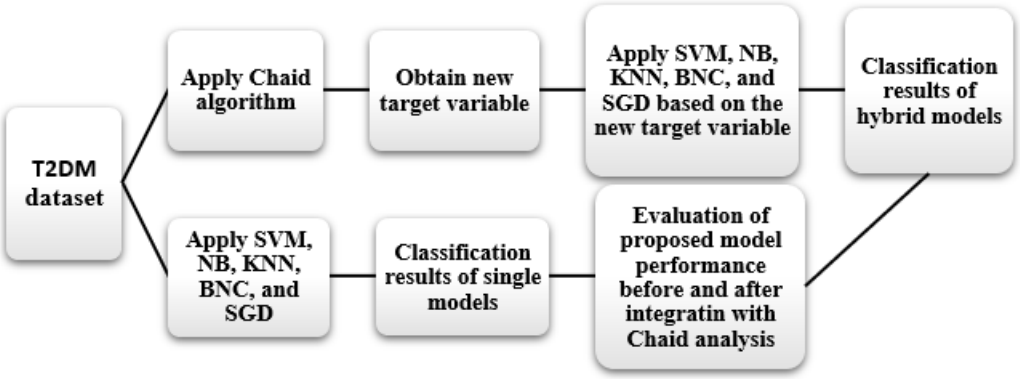

Figure 1: Implementation flowchart for proposed approach

3.2 Validation mechanism

A 10-fold cross-validation approach has been used to provide dependable and repeatable assessment of performance(Haq et al., 2020). The data was divided into 10 folds; for each fold, 9-folds are used for training, and one-fold is used as the test set. This is repeated with every fold in the dataset making up a test set, and the averaged results provide good, robust, and generalizable performance estimation. This procedure is detailed in Section 5 (Experimental Setup and Hyperparameters).

3.3 Experimental Setup and Hyperparameter Configuration

In order to uphold methodological rigor and avoid data leakage, CHAID was nested within each fold in the cross-validation process. The feature engineering and the labels were generated independently within the training partition, so no information from the test set influenced model development. The hybrid models were trained using the labels created from CHAID (793 T2DM, 93 Prediabetic) and assessed against the original clinical diagnoses (765 T2DM, 121 Prediabetic). This structure honors best practices in clinical predictive modeling, enabling a worst-case assessment while increasing the accuracy and generalizability of the model.
Each classifier was trained with the following hyperparameters:
SVM: RBF kernel, C=1.0 (regularization parameter), gamma=scale.
KNN: Euclidean distance for distance measure, uniform weighting for neighbors, k=5 neighbors.
Naïve Bayes: Gaussian and Laplace smoothed: $\alpha = 1.0$.
Decision Trees: Gini impurity (branching decisions), max depth = 5.
Random Forests: 100 trees, max depth = 10, max features = √.
Hyperparameters were selected with the aim of minimizing overfitting as much as possible while balancing the bias–variance.

3.4 Exploratory and Post-model Analyses (Bivariate analysis & Variable importance)

Alongside the primary predictive modeling, exploratory analyses of variables associated with T2DM were conducted. The Phi coefficient and Cramér's V were applied for binary and ordinal variables, respectively (Cohen, 2013) Variable importance was assessed using CHAID-derived normalized importance and permutation-based measures (Breiman, 2001) providing model-agnostic estimates of each predictor's contribution.

3.5 Performance evaluation metrics

The metrics used to assess the performance of the models were accuracy, sensitivity, specificity, and confusion matrices; these metrics contribute to a thorough assessment of classification efficiency in identifying Prediabetic cases vs. T2DM cases. The confusion matrix compares the observed and predicted classes. It has four outcomes: TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives). In our case, TP and TN represent a

situation in which the classifier correctly classified positive and negative ("T2DM " and "prediabetes") instances. The number of cases where the model misclassified positive as negative is denoted by FP, whereas FN represents patients' negatives that were incorrectly classified as positive. The confusion matrix is illustrated in Table 2.

Table 2: Confusion Matrix

	Positive (1)	Negative (0)
Positive (1)	(TP)	(FP)
Negative (0)	(FN)	(TN)

To reduce redundancy, a classifier should be evaluated using a set of performance metrics that are not closely related; these metrics are obtained by using a confusion matrix. In this paper, seven commonly evaluated metrics of two types are used: threshold metrics (precision, accuracy, sensitivity, specificity, F-score, and kappa) and rank metrics (ROC-AUC). The metrics are defined further below (Osi et al., 2020).

I. Accuracy is a widely used metric for monitoring the performance of machine learning models and is defined as the ratio of correctly classified instances to the total number of instances; however, for unbalanced data, this metric can be misleading.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TN + FP + FN + TP}} \quad (8)$$

II. F-score is the harmonic means of recall and precision is the F-score. The F-score is a number between 0 and 1; 1 means perfect.

$$\mathbf{F-Score} = \frac{\mathbf{2 * TP}}{\mathbf{(2 * TN) + FP + FN}} = \frac{\mathbf{2 * Recall * Precision}}{\mathbf{Recall + Precision}} \quad (8)$$

III. Kappa is a coefficient created to measure observed agreement normalized to chance agreement.

$$\boldsymbol{K} = \frac{\boldsymbol{P_{obs} - P_{chance}}}{\mathbf{1 - P_{chance}}} \quad (9)$$

IV. Sensitivity/recall is defined as the proportion of positive instances correctly classified as positive when compared to the entire positive set.

$$\text{Sensitivity/Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (10)$$

V. Specificity is defined as the proportion of negative instances correctly identified as negative out of the total observed negative instances.

$$\text{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN+FP}} \quad (11)$$

VI. Precision is the true positive divided by the sum of all positive predictions.

$$\text{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \quad (12)$$

VII. Receiver Operating Characteristics (ROC-AUC) is a graphical plot that depicts binary classifier performance for various thresholds. The true positive rate (TPR) is plotted against the false positive rate (FPR) at various threshold settings to generate the ROC. The area under the ROC curve (AUC) is a single value that ranges between 0 and 1 that summarizes the ROC performance. A higher ROC-AUC value indicates a classifier's superiority, and vice versa.

VIII. Time taken to build the model is the execution time representative of the total time taken for computations, including data splitting, data preprocessing, and model evaluation.

4. Application to T2DM data.

4.1 Data description

The population of the study is composed of T2DM patient records obtained from several Egyptian public and private hospitals between 2018 and 2023. To minimize sampling error, all available observations from the population were selected to provide an 886-observation sample size with seventeen independent variables and no missing values, as shown in Table 3.

Table 3: Variables description.

Variables	Description
Age (years)	20 to35=1, >35 to 60=2, >60=3
Gender	Female:0, Male=1

Residence	Rural =0, Urban =1
The level of obesity	<25 =1, 25 to 30 =2, >30 =3
High blood pressure	No=0, Yes=1
High levels of Triglycerides (mmol/L)	No=0, Yes=1
High levels of LDL (low-density lipoprotein) cholesterol	No=0, Yes=1
Low levels of Cholesterol HDL (mmol/L)	No=0, Yes=1
Smoking	No=0, Yes=1
Practicing physical activity	No=0, Yes=1
High levels of Serum creatinine	No=0, Yes=1
High levels of the hemoglobin A1c	No=0, Yes=1
Insulin resistance	No=0, Yes=1
Kidney disease	No=0, Yes=1
High levels of Uric acid	No=0, Yes=1
Family history of type 2 diabetes	No=0, Yes=1
Cardiac Diseases	No=0, Yes=1
Target variable (type 2 diabetics)	Prediabetics =0, Type 2 Diabetics =1

4.2Examination of Relationships between Explanatory Variables

The explanatory variables demonstrated tolerable levels of multicollinearity (all VIF < 10.0; High levels of the hemoglobin A1c = 1.89, insulin resistance = 1.76, High levels of Triglycerides (mmol/L) = 1.65) so they could be included collectively in the models.

With respect to the binary variables, there were only small (Φ = 0.267 for insulin resistance ↔ High levels of Triglycerides (mmol/L)), moderate (Φ = 0.423 for High blood pressure ↔ cardiac disease) and moderately strong (Φ = 0.456 for High levels of the hemoglobin A1c ↔ insulin resistance; Φ = 0.567 for High levels of Serum creatinine) associations. The only negative (inverse) but moderate association found was between smoking and physical activity (Φ = -0.345; i.e., an inverse lifestyle).

With regard to the ordinal variables, Cramér's V values suggested the expected monotonic trends for both small-to-moderate associations (age ↔ obesity; V = 0.298) and moderate associations (age ↔ High blood pressure; V = 0.398). All considered, there are valuable clusters across the sets of metabolic, cardiovascular, and lifestyle variables in this analysis with satisfactory redundancy, which provides both statistical and clinical utility.

4.3 preliminary stage

There is a preliminary stage before building the hybrid models, which is the stage of applying the CHAID algorithm to T2DM data to obtain the predicted value and then using it as an input (dependent variable) instead of using the raw data in the single models, with the goal of improving the classification performance of T2DM data. The frequency of T2DM data used in single models and hybrid models is shown in Table 4.

Table 4: The frequency of T2DM data

	Raw data Used in single models	Predicted Value as an output of CHAID algorithm used in hybrid models as input
	Frequency	Frequency
Prediabetics	121	93
T2DM	765	793

4.4 Confusion Matrix findings

Table 5 shows the confusion matrices of the single and hybrid models, illustrating an apparent performance difference between both dimensionality reduction approaches. Single classifiers resulted in large amounts of misclassification, indicating apparent confusion between prediabetic samples compared to T2DM samples, which was expected considering the raw data had a high number of features resulting in difficulty separating the classes. Hybrid models with CHAID reduced the amount of misclassification to a great extent. Specifically, SVM + CHAID and SGD + CHAID were able to completely separate the classes without any misclassification error (0 false positives and 0 false

negatives). The hybrid indicated that CHAID had taken the original set of data from the unprocessed state to a new set that also reduced class errors. Clearly the request for class separability is achieved partly through CHAID preprocessing, but more importantly through its function of transforming the overall set of data, thus enhancing sensitivity and specificity through dimensionality reduction. (Yang, Yi, Deng, & Sun, 2023).

Table 5 Confusion Matrices of Single and Hybrid Models

	Model	TP	FP	FN	TN	Total
Single Models	SVM	678	17	87	104	886
	Naive Bayes	662	17	103	104	886
	KNN	658	17	107	104	886
	BNC	662	17	103	104	886
	SGD	683	16	82	105	886
Hybrid Models (via CHAID)						
	SVM + CHAID	765	0	0	121	886
	NB + CHAID	764	0	1	121	886
	KNN + CHAID	728	4	37	117	886
	BNC + CHAID	762	0	3	121	886
	SGD + CHAID	765	0	0	121	886

Notes: Positive Class = T2DM (765 cases); Negative Class = Prediabetic (121 cases). Hybrid models were trained with CHAID-created labels (793 T2DM, 93 Prediabetic) and evaluated against recognized clinical diagnoses. This allowed Study to maintain methodological integrity and prevent data leakage.

4.5 Performance comparison between different models

Although Table 5 showed the confusion matrices of the predictive classifications, Table 6 detailed the performance outcomes. Accuracy, Precision, Sensitivity, Specificity, F-score, Cohen's Kappa, and ROC-AUC were calculated using 10-fold cross-validation using WEKA (versions 3.8.4/3.8.5). In all, the results supported that all hybrid models outperformed single models, using all performance measures. Both SVM + CHAID and SGD + CHAID had perfect ratings (100%) performance across all metrics, while all other hybrid models had results approaching perfection. The results demonstrate the value of using CHAID preprocessing, which increases performance of classifiers, decreases the rates of misclassifications and strengthens the clinical use of predictive models. The discussion below interprets the findings in relation to previous studies and considers how the findings could help develop useful clinical decision-support tools.

Table 6 Performance metrics of all the algorithms before and after integrating with CHAID algorithm.

	Classifiers	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F-score (%)	Kappa (%)	ROC-AUC (%)	Time taken (in seconds) to build the model
Single models	SVM	88.26	97.55	88.63	85.95	92.88	59.97	87.29	0.09
	NB	86.46	97.50	86.54	85.95	91.69	55.81	86.25	0.00
	KNN	86.00	97.48	86.01	85.95	91.39	54.96	85.98	0.00
	BNC	86.46	97.50	86.54	85.95	91.69	55.81	86.25	0.00
	SGD	88.94	97.71	89.28	86.78	93.31	61.86	88.03	0.04
Hybrid models	SVM via CHAID	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.03
	NB via CHAID	99.89	100.00	99.87	100.00	99.94	99.53	99.94	0.00
	KNN via CHAID	95.37	99.45	95.16	96.69	97.26	82.30	95.93	0.00
	BNC via CHAID	99.66	100.00	99.61	100.00	99.80	98.50	99.81	0.00
	SGD via CHAID	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.03

Table 5, Figure. 2, Figure. 3, and Figure. 4 show a comparison of the results obtained from single models (SVM, NB, KNN, BNC, and SGD) and hybrid models (SVM via CHAID, NB via CHAID, KNN via CHAID, BNC via CHAID, and SGD via CHAID). The highest accuracy of 100% is obtained by the proposed models of SVM via CHAID and SGD via CHAID compared with other models. In terms of precision, sensitivity, specificity, F-score, kappa, and ROC-AUC, the two models have the highest percentage values of 100, 100, 100, 100, and 100, respectively, when compared to other models. Hence, the accuracy of this classifier outperforms the others. Nevertheless, the two models have the second-highest value of the time taken to build the model, with a value of 0.03 minutes.

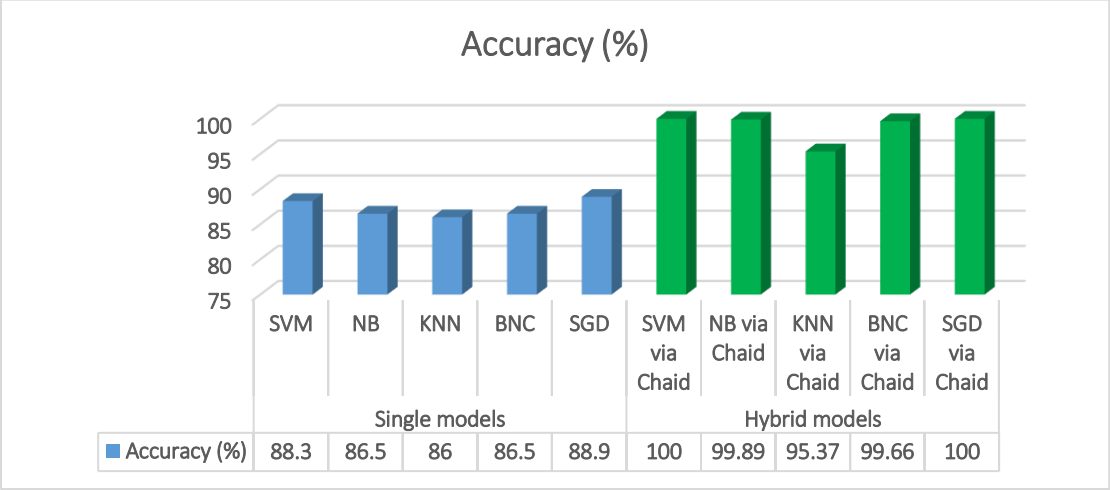

Figure 2: Model accuracy comparison graph

The accuracy of all single models improves when they are integrated with CHAID algorithm, as illustrated in Figure. 2.

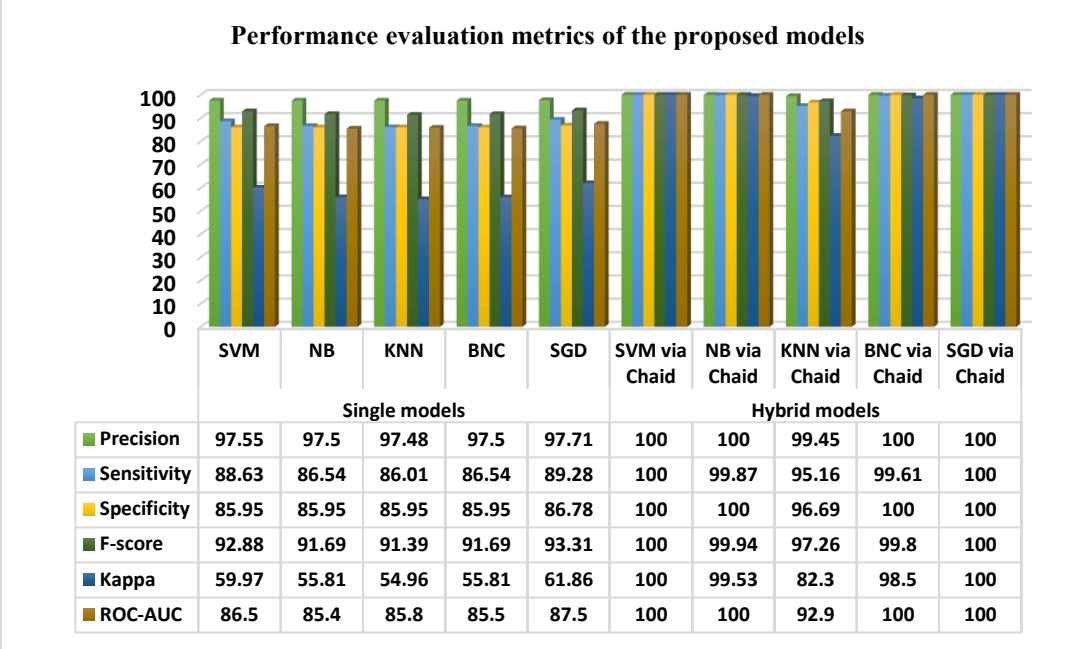

Figure 3: Graphical representation of performance evaluation metrics for the single models and the hybrid models

All metrics of all single models improve when they are integrated with the CHAID algorithm, as illustrated in Figure. 3

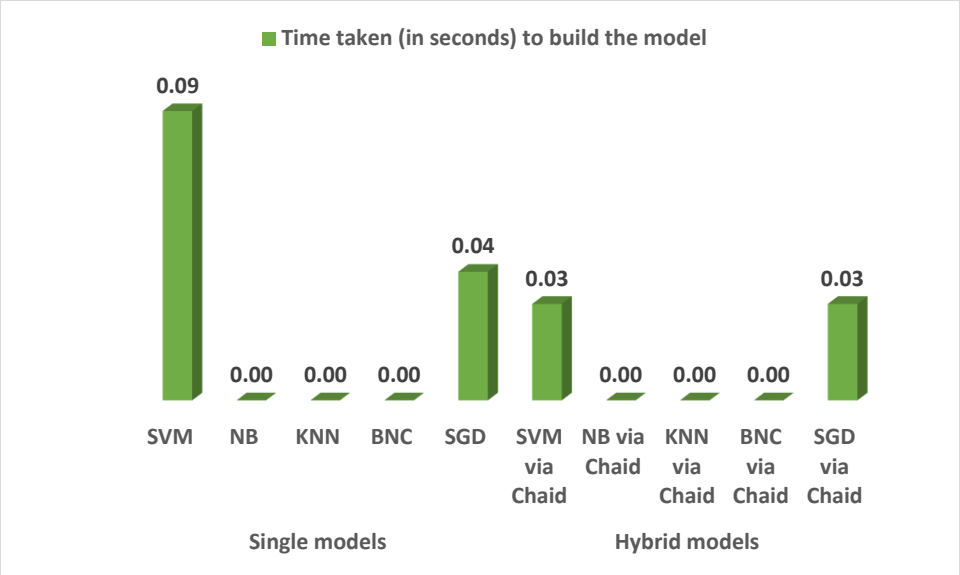

Figure 4: Performance evaluation graph for the time taken to build the single models and hybrid models.

Figure. 4 depicts a reduction in the time required to build the hybrid model. All previous results shown demonstrate how important integrating machine learning classifiers is for improving classification performance.

4.6 Ranking Importance of Variables

CHAID analysis, which uses the Chi-square statistics, reported the following variable-dependent hierarchy of importance for Type 2 Diabetes Mellitus (T2DM). The highest level of hemoglobin A1c (normalized importance = 1.000) as well as being an insulin resistant individual (0.892) were both the most intense predictors as clearly understood here as some form of (glycemic) dysregulation that may predispose a patient to T2DM, and that these forms of hemoglobin A1c and being an insulin resistant patient are also correlated themselves.

With the lipid-related variables, the highest association was with high levels of Triglycerides (0.756). Low levels of Cholesterol HDL (0.467) and High levels of LDL cholesterol (0.378) were both also validated as less strongly correlated but still fairly meaningful risk factors. This analysis also highlights the strong effects of age (years) (0.698), obesity (0.634), high blood pressure (0.587), and family history of type 2 diabetes (0.523). Lifestyle factors, and therefore "modifiable" factors, such as physical activity (0.412) and smoking (0.289), along with biomarkers such as high levels of Uric acid (0.334), and high Serum creatinine levels (0.245), were all substantial factors. The comorbidity of cardiac diseases (0.201) and kidney disease (0.167) appear to function more as associated co-factors than predictors and the demographic categories of gender (0.123) and residence (0.089) had the least relative impact on the classification.

In this way, the order of importance describes the prioritization of the core metabolic dysfunction, supported by demographic and genetic factors as opposed to downstream sequelae and demographic factors.

4.7 Risk of Overfitting and Model Robustness

While a 100% performance may indicate possible overfitting, there is nothing to suggest one is at risk given the researcher employed 10-fold cross-validation to evaluate the models on distinct partitions of the data. Furthermore, the segmentation derived from CHAID helps with generalization by capturing statistically significant signals rather than noise. The other evaluation metrics (i.e., precision, recall, F1, sensitivity, specificity) continuously verified the strength of our findings. Regularization parameters (i.e., penalty term C for SVM, a limitation on tree depth) were also used to avoid overly complex models.

5. Discussion and conclusion

The statistical analysis clearly indicated that high hemoglobin A1c and insulin resistance had the highest degree of association as predominant risk factors for T2DM, followed by dyslipidemia (high triglycerides, high LDL and low HDL), age, obesity, and High blood pressure. These findings provide a clinical evidence base for prioritizing key biomarkers when performing their diabetes screening. Regular testing of these parameters provides an opportunity for early identification of individuals at high risk of developing significant illness, providing opportunity for intervention employing lifestyle modification, commencing medical therapies and implementation of individualized treatment plans that can delay or prevent onset of complications due to diabetes.

From a modeling viewpoint, the hybrid classifiers that combined CHAID with SVM and SGD provided perfect discrimination in our sample without false positive or false negative rates. Clinically, this means these methods can give physicians far more diagnostic certainty than single models alone. Improved sensitivity means fewer missed cases of T2DM and earlier intervention, while improved specificity reduces unnecessary follow up tests and treatment. All of these direct benefits encourage more efficient patient triage, more effective use of healthcare resources, and improved targeting of prevention initiatives in primary care.

In addition, the brief timeframes for building the models suggest they may be incorporated into electronic health records or decision support systems for scheme delivery which could ultimately enable their real time usage in screening programs. An important consideration is that the internal validation (10-fold cross validation and regularization) demonstrates the stability of these models while external validation will remain necessary prior to routine clinical use.

Taken together, the application of CHAID using machine learning classifiers is meaningful progress in turning statistical performance into medical practice. In this regard, the models can improve accuracy of prediction but begin to influence clinical decision making, risk stratification, and management of early disease. As a translational bridge between adequate rule-based statistical modelling and improvements in patient care, these CHAID supported approaches have the potential to enhance diabetes screening and prevention initiatives. These outcomes are especially beneficial to institutions in high burden settings, where wastes from missed or delayed diagnoses of T2DM and subsequent inefficient allocation of resources to risk factor management and prevention measures could have unintended consequence on T2DM and its long consequence on health.

Even when the model showed strong in-sample performance, these findings should be viewed as being reflective of some dataset-specific patterns rather than generalizing - we definitely need better external validation - across multiple independent, multi-center cohorts - with calibration, predictive values, decision-curve analysis and subgroup performance analysis to assess generalization, fairness and clinical trustworthiness.

It is both a limitation and a strength, we cannot assume generalization across any healthcare system or country to be valid across other healthcare systems, but we also provided a rare context-specific example from a high-burden setting. Given that the model performed well, it offers a good foundation for adaptation and validation in other populations at this stage.

Moving forward, we need to externally validate the model against other cohorts and identify opportunities for real-world implementation - in particular integration with electronic health records and clinical workflows with appropriate governance frameworks to ensure equity, trust and sustainability in clinical use.

In Conclusion this study highlights the significant role of metabolic factors in the etiology of T2DM and illustrates that hybrid models employing CHAID supplementation applied for classification purposes can yield substantial improvements in classification objectives. The model will facilitate earlier interventions, more risk stratifying, and personalized management using the reliable and hierarchical risk profile provided. The findings are particularly useful, especially in an area of facing a high burden and limited resources, such as Egypt. Ultimately, however, there must be commitment throughout organizations to perform external validation and appropriate, ethically based implementation to allow for translatability into practice.

The implications for practice are important. Clinically, high-risk patients should be screened for the top ranked predictors - namely High levels of the hemoglobin A1c, insulin resistance and High levels of Triglycerides (mmol/L). Public health strategies that can target or tailor screening strategies that include all ages may be the best use of limited resources. Defining the odds as hybrid CHAID models will further enhance technology within EHRs and enhance clinician training to ensure optimal use of the models. Study and policy work moving forward will need to consider multi-site validation/triangulation, the costs, and implications regarding the potential for the framework to work in real-world contexts. Finally, equity and neutrality will necessitate audits for systematic fairness and transparency to build trust to demonstrate consistency to diverse types of patient's subpopulations. These elements reflect ways to realize statistical advances into practice. Ultimately, this will benefit individuals who are susceptible to T2DM impacting positive health outcomes.

Statements and Declarations

This study has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal. There are no conflicts of interest to disclose.

Funding

For this work, I received no funding.

Competing Interests

There are no financial or non-financial interests in the work submitted for publication, either directly or indirectly.

Data availability statements

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

ORCID iDs

Walaa Elaraby **https://orcid.org/0000-0002-3804-5154**

1. References

2. Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC bioinformatics, 24(1), 337.
3. Anuradha, C., & Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. Indian Journal of Science and Technology, 8(15), 1-12.
4. Atiyah, O. S., & Thalij, S. H. (2022). Evaluation of COVID-19 Cases based on Classification Algorithms in Machine Learning. Webology, 19(1).
5. Avula, A., & Asha, A. (2018). Improving prediction accuracy using hybrid machine learning algorithm on medical datasets. IJSER, 9(10), 1461-1467.
6. Azami, G., & Ebrahimy, B. (2023). alarming prevalence of comorbid conditions in adults with type 2 diabetes in Iran. Journal of Diabetes & Metabolic Disorders, 22(2), 1809-1811.
7. Bharadwaj, A., & Minz, S. (2012). Hybrid approach for classification using support vector machine and decision tree. Paper presented at the Proceedings of the International Conference on Advances in Electronics, Electrical and Computer Science Engineering-EEC.
8. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
9. Çalişir, D., & Doğantekin, E. J. E. S. w. A. (2011). An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. 38(7), 8311-8315.
10. Cohen, J. (2013). Statistical power analysis for the behavioral sciences: routledge.
11. Doppala, B. P., Bhattacharyya, D., Chakkravarthy, M., Kim, T.-h. J. D., & Databases, P. (2021). A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. 1-20.
12. Farag, H. F. M., Elrewany, E., Abdel-Aziz, B. F., & Sultan, E. A. (2023). Prevalence and predictors of undiagnosed type 2 diabetes and pre-diabetes among adult Egyptians: a community-based survey. BMC Public Health, 23(1), 949.
13. Gashti, M. Z. J. I. J. o. A., & Sciences, A. (2017). A novel hybrid support vector machine with decision tree for data classification. 4(9), 138-143.
14. Grwany, T. J. T. E. J. o. L. E. (2021). Performance evaluation in Arabic sentiment analysis during the COVID-19 pandemic. 8(2), 16-27.
15. Haq, A. U., Li, J. P., Khan, J., Memon, M. H., Nazir, S., Ahmad, S., . . . Ali, A. (2020). Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. Sensors, 20(9), 2649.
16. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. J. M. I. S. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. 2018.
17. Hassan Yaseen, H., & Alibraheemi, K. J. N. (2022). Classification Covid-19 disease based on CNN and Hybrid Models. 8039-8054.
18. Hegazi, R., El-Gamal, M., Abdel-Hady, N., & Hamdy, O. (2015). Epidemiology of and risk factors for type 2 diabetes in Egypt. Annals of global health, 81(6), 814-820.
19. Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Journal of the royal statistical society: series c (Applied statistics), 29(2), 119-127.

20. Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. J. B. e. d. (2019). Predictive models for diabetes mellitus using machine learning techniques. 19(1), 1-9.
21. Li, G., Chen, K., Yang, H. J. P. S., & Protection, E. (2022). A new hybrid prediction model of cumulative COVID-19 confirmed data. 157, 1-19.
22. Li, X., Zhang, J., & Safara, F. (2023). Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. Neural processing letters, 55(1), 153-169.
23. Mowafy, M. A. A. S., Shallan, W. M. E. M. J. R. o. E., & Science, P. (2021). Improving the classification accuracy using hybrid techniques.
24. Murthy, N. S., & Arunadevi, B. J. S. M. i. M. R. (2021). An effective technique for diabetic retinopathy using hybrid machine learning technique. 30(4), 1042-1056.
25. Mutlu, G., & Acı, Ç. İ. J. P. C. (2022). SVM-SMO-SGD: A hybrid-parallel support vector machine algorithm using sequential minimal optimization with stochastic gradient descent. 113, 102955.
26. Nikam, S. S. J. O. J. o. C. S., & Technology. (2015). A comparative study of classification techniques in data mining algorithms. 8(1), 13-19.
27. Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., . . . Hagins, H. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet, 402(10397), 203-234.
28. Osi, A. A., Abdu, M., Muhammad, U., Ibrahim, A., Isma'il, L. A., Suleiman, A. A., . . . Ringim, M. Z. (2020). A Classification Approach for Predicting COVID-19 Patient's Survival Outcome with Machine Learning Techniques. medRxiv.
29. Rokach, L. (2010). Ensemble-based classifiers. Artificial intelligence review, 33(1), 1-39.
30. Saxena, R. (2021). Role of K-nearest neighbour in detection of Diabetes Mellitus. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 373-376.
31. Sheela, M. S., & Arun, C. J. I. J. o. I. T. (2022). Hybrid PSO–SVM algorithm for Covid-19 screening and quantification. 1-8.
32. Sneha, N., & Gangil, T. J. J. o. B. d. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. 6(1), 1-19.
33. Wang, Y., Ni, X. S., & Stone, B. (2018). An automatic interaction detection hybrid model for bankcard response classification. Paper presented at the 2018 5th International Conference on Systems and Informatics (ICSAI).
34. Yang, Y., Yi, F., Deng, C., & Sun, G. (2023). Performance analysis of the CHAID algorithm for accuracy. Mathematics, 11(11), 2558.
35. Yoon, E.-J. (2012). Improvement of the securing RFID systems conforming to EPC class 1 generation 2 standard. Expert Systems with Applications, 39(1), 1589-1594.
36. Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists: " O'Reilly Media, Inc.".