Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Critical Factor Analysis for prediction of Diabetes Mellitus using an Inclusive Feature Selection Strategy

E. Sreehari & L. D. Dhinesh Babu

To cite this article: E. Sreehari & L. D. Dhinesh Babu (2024) Critical Factor Analysis for prediction of Diabetes Mellitus using an Inclusive Feature Selection Strategy, Applied Artificial Intelligence, 38:1, 2331919, DOI: 10.1080/08839514.2024.2331919

To link to this article: https://doi.org/10.1080/08839514.2024.2331919

Published online: 01 Apr 2024.

APPLIED ARTIFICIAL INTELLIGENCE
2024, VOL. 38, NO. 1, e2331919 (31 pages)
https://doi.org/10.1080/08839514.2024.2331919

Taylor & Francis
Taylor & Francis Group

Critical Factor Analysis for prediction of Diabetes Mellitus using an Inclusive Feature Selection Strategy

E. Sreehari and L. D. Dhinesh Babu

School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, India

ABSTRACT

Diabetes mellitus is a metabolic disorder that significantly implicates serious consequences in various parts of the human body, such as the Eye, Heart, kidney, Nerves, Foot, etc. The identification of consistent features significantly helps us to assess their impact on various organs of the human body and prevent further damage when detected at an early stage. The selection of appropriate features in the data set has potential benefits such as accuracy, minimizing complexity in terms of storage, computation, and positive decision-making. The left features might contain potential information that would be useful for analysis. In order to do effective analysis, additionally, all features should be studied and analyzed in plausible ways, such as using more feature selection (FS) methods with and without standardization. This article focuses on analyzing the critical factors of diabetes by using univariate, wrapper, and brute force FS techniques. To identify critical features, we used info gain, chi-square, RFE, and correlation using the NIDDK data. Later, distinct machine learning models were applied to both phases of the feature sets. This study was carried out in two phases to evaluate the efficacy of the techniques employed. The performance has been assessed using accuracy, F1score, and recall metrics.

ARTICLE HISTORY
Received 19 August 2023
Revised 1 February 2024
Accepted 9 March 2024

Introduction

Globally diabetes is recognized as a exacerbate disease and is listed as the 9th most influenced disease causing more mortality rates. According to World Health Organization (WHO) diabetes had caused approximately 1.5 million deaths directly in the year 2019. The International Diabetes Federation (IDF) reported in the year 2021 that globally more than 463 million people were affected with diabetes mellitus between the age group of 20–79 years. And it was estimated that by the year 2045 the diabetes mellitus count may reach an alarming level of 700 million. In addition, IDF stated that next to China,

CONTACT L. D. Dhinesh Babu lddhineshbabu@vit.ac.in School of Computer Science Engineering and Information Systems (SCORE), Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India

India has the highest number of diabetic patients which would be approximately 77 million people (Statista 2023). Insulin regulates the blood sugar levels in order to ensure the consistent functioning of the human body organs in a normal manner. When there is no sufficient secretion of insulin hormone by the pancreas, This leads to Hyperglycemia (raised blood sugar). The uncontrollable presence of excess blood glucose levels for a longer period of time is often referred as Diabetes (Gromova, Fetissov, and Gruzdkov 2021).

The longer diabetes mellitus leads to serious complications such as kidney failure, blindness, and heart strokes, foot ulcer infections, hearing impairment, more time to heal wounds, and the Alzheimer's disease etc (Alam et al. 2021; Papatheodorou et al. 2015; Tomic, Shaw, and Magliano 2022; Unnikrishnan, Anjana, and Mohan 2016). The various reasons for raise in glucose levels are due to stress, lack of proper dietary habits and sleep, genetic mutations and improper functioning of pancreas (CDC 2023).

A study by ICMR and MDRF reveals that 25 million people are found to be prediabetes stage. There is approximately 10% of the population India with 74 million confirmed diabetes reported in the year 2021. According to IDF Diabetes Atlas 2021, the Diabetes prevalence has reached globally with over half a billion confirmed diabetics in 2021 as shown in Figure 1 (IDF Diabetes Atlas 2023). Figure 1 displays an exponential behavioral trend that reflects the enormous growth rate of people affected by diabetes over the past two decades. Hayden E. Klein's study showed that the numbers are projected to double within a short period of time (Diabetes Prevalence Expected to Double Globally by 2050 2023). To effectively address this issue with a major concern, it is

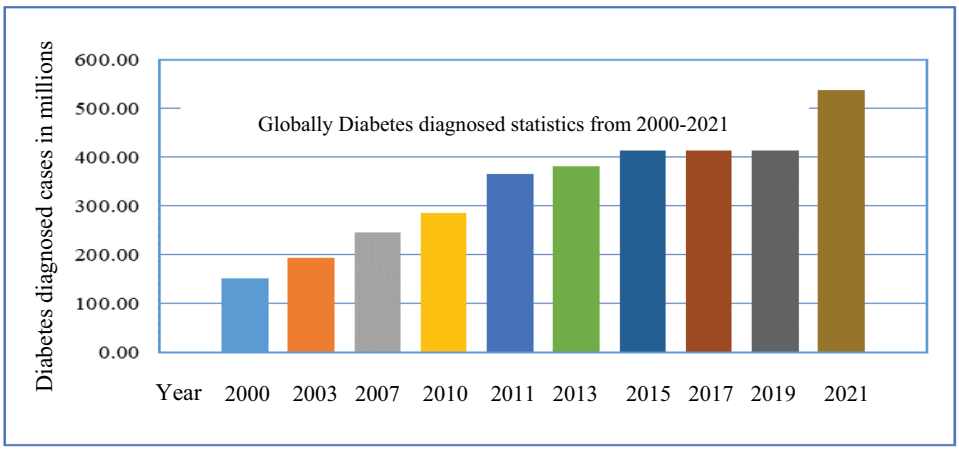

Figure 1. Diabetes mellitus impact statistics according to IDF report (source: (IDF Diabetes Atlas 2023)).

imperative to actively identify and analyze the crucial features of diabetes. Therefore, to control the diabetes, enough awareness is required as well as identifying the crucial factors causing diabetes and monitoring them would assist the people to maintain organs free from its impact.

In simple terms, the prevention and management of diabetes can help to reduce diabetes complications in order to improve the quality of life for people. The affected patients must have an effective imperative analysis to keep diabetes under control and to provide an enhanced treatment. Analyzing the diabetes critical features using multi criteria feature selection with dual data analysis consideration with and without standardization would help us to address this issue. Feature selection applications are enormous which would be used for decision making in distinct fields such as e-commerce, business, government services, graph theory, metal industries, medical applications, disease-oriented prediction and treatment, gene selection, Microarray data etc. As shown in Table 1 the Feature selection process involves a series of steps, such as search direction, search strategy, evaluation strategy, stopping criteria, and validation process (Sahu, Dehuri, and Jagadev 2018). In order to do effective feature selection, features are classified as redundant, noisy, weakly as well as strongly relevant and irrelevant features in general (Yu and Liu 2004). The right feature selection helps to run the model more efficiently, reduces computational and storage costs, and boosts processing power (Gutkin, Shamir, and Dror 2009). Features are classified based on noise, relevancy, redundancy, and inconsistency properties. The measures such as distance based, correlation based (Buyrukoğlu and Akbaş 2022), consistency and threshold methods, etc., are used for feature analysis processing (Aha et al. 1991; Kira and Rendell 1992). To ensure safety, reliability, and to accelerate the battery development cycle Zi Cheng Fei (Fei et al. 2021) proposed an early lifetime prediction for a battery with a combination of a Machine Learning (ML) model and a wrapper feature selection approach. He manually crafted 42 features based on the first-100-cycle charge-discharge raw data. Numerical experiments and paired t-tests are conducted to statistically evaluate the performance of the proposed framework. The support vector machine (SVM) model combined with wrapper feature selection presents the best result for battery lifetime prediction. In industry informatics privacy preserving for industrial applications is crucial issue and to address this issue, the authors Tao Zhang et al. (Zhang et al. 2020) introduced a correlation reduction scheme with differential private feature selection by considering the issue of privacy loss while data has a correlation in ML tasks which may lead to more privacy leakage. Experiments show that the proposed scheme obtained better prediction results with ML tasks and fewer mean square errors for data queries compared to existing schemes.

Table 1. The core emphasizing relative concepts required for feature selection exploration.

Name	Definition	Method
Search direction	A starting tactic to carry out the process.	Forward, backward, Random (Ang et al. 2016)
Search strategy	Way of accessing or selecting features	Exhaustive, sequential, Randomized (Venkatesh and Anuradha 2019)
Evaluation strategy	Evaluation states that to draw a conclusion based on the features Merit/significance/worth, standard	Filter (Bommert et al. 2020), wrapper, Hybrid, Embedded, Heuristic and meta heuristic (Alyasiri et al. 2022; Hsu, Hsieh, and Da Lu 2011; Hu and Wu 2010; Pirgazi et al. 2019)
Stopping criteria	Stopping of the feature selection process by using some constraints	Fixed feature count, number of iterations, evaluation function-based iterations (Sahu, Dehuri, and Jagadev 2018)
Validation process	The resultant feature set is evaluated in validation process	Cross validation (Browne 2000), confusion matrix (Kulkarni, Chong, and Batarseh 2020), Jaccard and other metrics (Dalianis 2018) etc.

Kushan De Silva et al (De Silva, Jönsson, and Demmer 2020) demonstrated the value of combining feature selection with ML to identify the predictors that could be useful to enhance prediabetes prediction and clinical decision-making. The authors analyzed a sample of 6346 men and women enrolled in the National Health and nutrition examination survey 2013–2014. Four machine learning algorithms were applied to 46 exposure variables in original and resampled training datasets built using 4 resampling methods. A range of predictors of prediabetes was identified and the result of prediabetes prevalence was 23.43%. Shivani Jain et al. (Jain and Saha 2022) developed the rank-based univariate parameter selection strategy using ML classifiers to detect the code smells for vast and sophisticated software. Mutual information (MI), fisher score, and univariate ROC – AUC feature selection techniques were used with brute force and random forest correlation strategies.

The authors compared and analyzed the classifiers' performance with and without feature selection. Rung Ching Chen et al. (Chen et al. 2020) introduced a new feature selection algorithm strategy for data classification based on ML methods. The authors of the article used three popular datasets such as bank marketing, car evaluation database and human activity recognition using smartphones for doing experimentation. Tadist et al. (Tadist et al. 2019) highlights four main reasons why feature selection is essential. In order to reduce the complexity of genomic data analysis and to get useful information quickly. To overcome the limitation of lower performance due to filter methods Yosef Masoudi et al. (Masoudi-Sobhanzadeh, Motieghader, and Masoudi-Nejad 2019) developed the Featureselect software application. This application was developed by using ten optimization algorithms along with filter methods and three learners. The data sets such as Carcinoma, Drive, Air, Drug, Social, and Energy were tested with the developed application. Out of all methods World Competitive Contests (WCC), League Championship Algorithm (LCA), Forest Optimization Algorithm (FOA), and Learning Automata (LA)

performed well and the results showed that wrapper methods are better than the filter methods. Yue Liu et al. (Liu et al. 2020) introduced the field experts-based feature selection weighted score model for material manufacturing properties target identification. The method Data-driven Multi-Layer Feature Selection (DMLFS) were implemented with 7 material experts for features consideration, max-mini normalization and ML models for identifying the proper descriptors of material. He tested his proposed methodology on ten material-related data sets and finally proved that the proposed method able to work effectively to identify the targeted properties of materials by ensuring good accuracy.

The early prediction of diabetes levels is more imperative when affected by it. In order to keep the diabetes under control and to provide enhanced treatment process, diabetes critical factors analysis would help the affected patients to maintain their health effectively as well as to avoid diabetes complications. Accomplishing this issue would require a potential study over the diabetes critical features. The importance of this study is to analyze the impact of globally prompted diabetes mellitus and the essence to work toward diagnosing, controlling and preventing diabetes. So, we carried out our work to assess the DM vital features analysis in a significant manner by proposing the work in a two-phase manner with and without standardization. The concepts of feature selection and balancing are the major key strategies involved for doing this analysis by applying machine learning algorithms at the end of both phases.

The study has implemented and incorporated its results into various sections of the article while achieving the following objectives.

(i) The study has been explored to identify the diabetes mellitus common root causes, its implications over other body parts of the diabetes affected people. In addition, the diabetes statistics from 2000 to 2021 are explored to show the stature of diabetes impact globally over the years.
(ii) Identifying the critical factors of diabetes in a standardized and non-standardized way by using multi criteria-based feature selection methods.
(iii) To compute and analyze the consistency, critical behavior of the feature selection methods based on dual nature of data by analyzing the individually obtained feature sets from both the phases using different classification models.
(iv) To compare and analyze the results of the ML procedures employed over the diabetes data by using the distinct metrics such as accuracy, precision, recall, F1-score, sensitivity and specificity.

The article presents an outline in the following manner: In Section 2, we have discussed the various authors' works on diabetes mellitus analysis and feature selection concepts, including their limitations. In Section 3, we have talked about the system architecture, feature selection methods applied, the data set, and the coding organization process as well. We discussed the experimental results in Section 4. In subsequent sections 5 and 6, we described the conclusion and future work, respectively.

Literature Survey

The preliminary study has identified that diabetes is primarily caused by risk factors such as lifestyle, hereditary factors, psychosocial elements, demographic features, family history, ethnicity, and certain medical conditions. This section covers a comprehensive review of various diabetic analysis-related strategies. From 2019 to 2023, we actively explored a holistic view of critical analysis for diabetes with greater emphasis. We identified gaps and sought to understand the depth of knowledge on diabetes mellitus by articulating pertinent works during this time period. The works were studied, analyzed and discussed based on two primary requirements they were feature selection conceptual methods and the diabetes critical analysis. Subhash and Goel (Gupta and Goel 2023) presented an ML model for predicting diabetes in patients. Their study compares different classification algorithms and improved their performance by preprocessing and tuning hyperparameters. The RF model achieved an F1 score of 75.68% and an accuracy of 88.61% on the PIMA dataset. Jahan and Hoque (Kakoly, Hoque, and Hasan 2023) proposed the concept with PCA, IG two folded-based parameter selection for predicting the diabetes risk factors using ML algorithms (DT, RF, SVM, KNN LR). The primary data used in the study were collected based on the Helsinki Declaration, 2013, out of which 738 records were included in the final analysis. The study achieved an accuracy level of over 82.2%, with an AUC value of 87.2%. Saxena et al (Saxena et al. 2022). aimed for a comparative study of classifiers and feature selection methods for accurate prediction of diabetes. The study used four classifiers (DT, KNN, RF, MLP) and three FS techniques (IG, PCA, SU) on the PIMA Indians diabetes dataset. The RF classifier achieved the best accuracy of 79.8%. Gupta et al. (Gupta et al. 2022) implemented an ML model for diagnosing diabetes early with greater accuracy. The hybrid model used NSGA-II and ensemble learning. The dataset used in the study comprises 23 features, with 1288 instances of patients. NSGA-II-XGB approach obtained the better result with an accuracy of 98.86%. Ayse Dogru et al. (Doğru, Buyrukoğlu, and Arı 2023) introduced a super-ensemble learning model that enables the early detection of diabetes mellitus. They developed an ensemble model using Grid Search, Chi square methods for hyperparameter tuning, and feature selection. The authors experimented using four

base-learners (LR, DT, RF, and gradient boosting) and a SVM meta-learner over three distinct data sets (diabetes risk prediction, PIMA, and 130 US hospital diabetes datasets). The proposed model effectively diagnosed diabetes with greater accuracy scores. Sabitha and devi (Sabitha and Durgadevi 2022) introduced a new framework that combined SMOTE augmentation, RFE-based FS, and preprocessing approaches in order to do diabetes prediction accurately. They experimented on the PIMA data set using the methods SVC, KNN, RF, LR, SVC and NB. The proposed method using RFE with RF regression FS achieved consecutive accuracy scores for RF, DT, and SVC of 81.25%, 81.16%, and 82.5%.

To determine the diabetes early Mehmet et al. (Gürsoy and Alkan 2022) investigated the diabetes which was collected from the Diabetes Specialization laboratories of Medical City Hospital and Al Kindy Training Hospital. The collected data were classified: they were normal, pre-diabetes and diabetes classes. The implementation was carried out by using the DL methods such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) with permutation feature importance for diabetes data analysis and diagnosis. Their created model demonstrated that the results would give medical practitioners a predictive tool for efficient decision-making that could help in the early diagnosis of the condition. Tiwari et al. (Tiwari and Singh 2021) proposed the methodology using attribute selection and classification methods predicted diabetes. The technique combined RFE, RF, and the Apriori methodology to identify the significant analysis of the diabetes variables. The XGBoost approach obtained better accuracy. An early and precise diabetes diagnosis approach was implemented by Amit and Chinmay (Kishor and Chakraborty 2021) by utilizing the correlation methodology and five ML classifiers (SVM, KNN, RF, LR, NB). They demonstrated that the RF classifier was capable of achieving higher scores for accuracy, sensitivity, specificity, and AUC (97.81, 99.32, 98.86, and 99.35, respectively). The authors Nagaraj et al. (Nagaraj et al. 2021) developed the system to classify diabetes-type diagnoses by considering the multiple patient features. Artificial flora-based FS and gradient-boosted tree (GBT) based classification methods were applied after processing the data format conversion and transformation. Their implemented methodology obtained better scores in terms of accuracy, recall, precision, and F-score. Oyebisi et al. (Oladimeji, Oladimeji, and Oladimeji 2021) showed the strategy by employing ranking-based feature selection techniques and ML algorithms (KNN, J48, NB, and RF) to detect diabetes at an early stage. It was established that their suggested model was more effective than research from the past. Their research highlights the value of ML in healthcare and the potential applications it offers for diagnosing and treating illnesses. A thorough review of research on developing precise models for anticipating complications of diabetes using ML approaches was presented by Anuradha et al.

(Madurapperumage, Wang, and Michael 2021). This study intended to aid model developers, researchers, and physicians in further exploration of diabetic study.

Bharath and Udaya Kumar (Chowdary and Kumar 2021) proposed a Neuro-Fuzzy model that used feature extraction to improve diabetes classification. The model was tested on the PIMA diabetic dataset and outperformed existing ML models. Jayroop et al. (Ramesh, Aburukba, and Sagahyroon 2021) established the remote-based monitoring framework system for effective diabetes control employing handheld, intelligent accessories, and individual medical devices. Using an SVM the authors developed a diabetes risk forecasting model. The suggested method achieved greater flexibility and vendor connectivity while enabling educated decisions based on recent diabetes risk projections and lifestyle insights. Authors Juneja et al. (Juneja et al., 2021) described the use of ML approaches and a multi-criteria decision-making framework for diabetes analysis. The Pima diabetes data were used in the research to investigate several predictive algorithms. The results show that supervised learning algorithms performed better than unsupervised ones and that the maintenance of an active lifestyle could be supported by a diabetes early diagnosis.

Jiaqi Hoq et a.l (Hou et al. 2020) introduced a method aimed to predict the prevalence of diabetes with the obtained physically examined parameters. The parameter selection methods RFE and F-score were employed subsequently DT, RF, LR, SVM, and MLP were applied to the selected features for diabetes prediction. However, certainly, the system achieved higher accuracy for RF and F-score combination. The developed hybrid classification model by Mishra et al. (Mishra et al. 2020) used the Adaptive based enhanced genetic algorithm of MLP for diagnosing diabetes. Their work obtained an accuracy of 97.8% approximately and execution time took 1.12 seconds with the help of attribute optimization technique. The model could assist medical experts to determine risk factors for type 2 diabetes. A Mendelian randomization study that identified the risk factors of diabetes type 2 was presented in the article. The study discovered proof of causal relationships between 34 exposures, including systolic blood pressure, depression, sleeplessness, and smoking.

Sneha and Gangil (Sneha and Gangil 2019) experimented with using ML for the early detection of diabetes by selecting significant features from the dataset. DT, RF algorithms achieved high specificity, and NB had the best accuracy. The research also aimed to improve classification accuracy by selecting optimal features. Choubey et al. (Choubey et al. 2020) developed a system for efficient diagnosis of diabetes using two phases. One phase involves dataset collection and analysis with KNN, LR, and DT methods. Phase 2 involved applying PCA and PSO algorithms to the data. The approach was more efficient in terms of computation time and accuracy and had the potential for early diagnosis of

other medical diseases. On adult demographic datasets Diwan (Alalwan 2019) formulated a technique for identifying type 2 diabetes by utilizing SOP and RF algorithms. The Self-Organizing Map method outperformed other algorithms in terms of accuracy. The author recommended that the system must be combined with the diabetic detecting equipment for quicker diagnosis. Simon Fong et al. (Li and Fong 2019) constructed a Coefficient of Variation (CV) parameter selection method to enhance diabetes accuracy. By utilizing the Prima diabetes dataset, the CV technique performed better than conventional FS methods because it excluded attributes with low data dispersion. The method came under the categorial behavior of a filter-based scheme which might take a small time for analysis but it did not guarantee obtaining greater accuracy.

Fisher's linear discriminant analysis (LDA) based concept systemized by Sheik and Selva (Sheik Abdullah and Selvakumar 2019) to identify risk factors for type II diabetes with the help of DT and PSO. The technique exhibited increased precision in detecting risk variables and could be applied to various chronic diseases. The system recognized the strong relationship among the MBG-PPG-A1c-FPG (Mean Blood Glucose – Postprandial Plasma Glucose – Glycosylated Hemoglobin – Fasting Plasma Glucose) factors. The potential benefits of using feature selection algorithms provided insights into the current state-of-the-art in this field. The works stated earlier were helpful for expanding the scope of research on feature selection algorithms. Added to that a new improved version of taxonomy for emerging development in multidisciplinary fields yielded more benefits.

The aforementioned works (Table 2) by different authors have implemented their experimental works with a limited number of selected features, using only one feature selection method, which cannot be deemed as a virtuous approach. In addition, none of the works have not discussed the consistency of the feature selection in respect of standardization before and after. Using mean or median values to replace missing data, removing corresponding records or columns, and employing other methods to handle missing data may introduce bias into the results. After extensively analyzing the various literature on diabetes mellitus, it has been determined that identifying and organizing the hierarchy of critical factors is crucial in understanding its importance. Conducting critical factors impact analysis is essential in order to effectively handle data pertaining to prior and posterior standardization. The article aims to improve the critical factors analysis of diabetes by utilizing a combination of data standardization before and after, in order to address key limitations such as inconsistencies in the quality and quantity of available data that may hinder result comparisons due to lack of standardization across studies.

Table 2. The relevant works of literature concerning diabetes were studied.

References	Proposed objective	Data methods used	Metrics used	Results	Limitations
(Chang et al. 2023)	ML based e-diagnosis system for diabetes analysis with supervised models	Pima diabetes data, NB, RF, J48 DT	accuracy, precision, sensitivity, and specificity	NB works well with selected features and RF performed well with all	The accuracy may be enhanced by using suitable pre-processing techniques
(Saxena et al. 2022)	To improve diabetes factors analysis using correlation	Pima diabetes data, MLP, DT, KNN, RF	Accuracy, classification error	RF obtained 79.8% accuracy	Few predictors considered to predict diabetes risk
(Chatrati et al. 2022)	Diabetes App to predict hypertension using glucose and BP	Pima diabetes dataset, SVM, KNN, LR, DT, LDA	Accuracy	SVM and LDA classification algorithms performed better	The author only focused on readings of patient's glucose and blood pressure
(Joshi and Dhakal 2021)	To analyze diabetes factors with LR and DT	Pima data set, LR, DT	Accuracy, cross validation error	LR obtained an accuracy of 78%	It would be interesting to investigate performance of the proposed method on large data
(Gupta and Goel 2023)	Diabetes analysis by Hyperparameter tuning	diabetes data, RF, SVM, KNN, LR, DT	Accuracy, AUC	Achieved an accuracy of 0.92, F1 score of 0.91, and AUC of 0.96	The study did not consider the effect of lifestyle factors on diabetes
(Kakoly, Hoque, and Hasan 2023)	Data-driven diabetes prediction using hybrid IG, PCA	Pima diabetes data, LR, DT, RF, SVM, KNN, NB	Accuracy, Sensitivity, Specificity, F1 Score	LR achieved an accuracy of 82% and AUC 87% with 5 features	The authors didn't focus on estimating computational complexity
(Gupta et al. 2022)	Meta-heuristic FS with XGBoost for diabetes prediction	F NSGA-II diabetes data, DT, SVM, NB, XGB, MLP, GA, PCA, PSO	Accuracy, precision, Sensitivity, Specificity, F1 Score	NSGA-II-XGB obtained best accuracy of 98.86%	Experimented on single hybrid diabetes data and suggested to work with multiple diabetes sets
(Doğru, Buyrukoğlu, and Arı 2023)	Early prediction of diabetes using hybrid model (chi-square, grid search)	stage diabetes risk prediction, PIMA,130-US hospitals dataset, LR, DT, RF, SVM meta learner	Accuracy, Sensitivity, Specificity, F1 Score, AUC	SVM obtained 99.6%, 92%, 98% accuracy levels over 3 datasets	Suggested to compare efficacy of ML and DL algorithms using diabetes data
(Sabitha & Durgadevi, Sabitha and Durgadevi 2022.)	Improving the diabetes diagnosis prediction with RFE	Pima diabetes data, LR, RF, SVM, KNN, NB, DT	Accuracy, F1 Score, AUC	With RFE and SMOTE RF (81%), DT (81%), SVC (82%) performed well	The authors proposed method not experimented with larger dataset and mixed data sets
(Tiwari and Singh 2021)	Diabetes disease prediction	Pima diabetes data, RFE with RF, XGB, ANN	Accuracy	XGBoost obtained better accuracy 78.91%	The proposed approach better to compare with multiple data sets

(*Continued*)

Table 2. (Continued).

References	Proposed objective	Data methods used	Metrics used	Results	Limitations
(Kishor and Chakraborty 2021)	Diabetes accurate prediction with correlation, SMOTE	Pima diabetes data, LR, SVM, RF, NB, KNN	Accuracy, Sensitivity, Specificity, AUC	RF gets 97% accuracy, 99.32% sensitivity, 98.86% specificity	The author suggested to work with other chronic disease data relevant to smart city based
(Oladimeji, Oladimeji, and Oladimeji 2021)	Likelihood diabetes prediction with Relief-F and correlation	Diabetes data from Sylhet Hospital of Sylhet, Bangladesh, KNN, J48, NB, RF	Accuracy, ROC AUC, PR AUC and F-measure	RF achieved with greater accuracy of 98.3%	Suggested to consider BMI, body size and height to analyze the role of parameters for diagnosing diabetes
(Mishra et al. 2020)	Diabetes analysis Enhanced adaptive hybrid model	Pima diabetes data, EAGA-MLP, NB, DT, SVM	Accuracy, Precision, F1 Score	MLP Achieved greater accuracy, precision and F-score.	The authors suggested to apply upgraded model over other disease disorders.
(Choubey et al. 2020)	Evaluation of classification methods with PCA and PSO for diabetes	Pima diabetes data, Bombay medical hall, Jharkhand, LR, ID3 and C4.5 DT, RF, NB, KNN	Accuracy, Sensitivity, Specificity, F1 Score, AUC	Achieved an accuracy of 0.85, sensitivity of 0.80, specificity of 0.90, F1 score of 0.82, AUC of 0.89	proposed approach computed with less computation time and increased accuracy.
(Alalwan 2019)	Diabetic analytics using data mining approaches in type 2 diabetes dataset	Adult population data set, Self-Organizing Map (SOM), RF, NB, DT, SVM, MLP	Accuracy	SOM performed better than other algorithms	does not provide a comprehensive analysis of the limitations and future scope of the proposed method
(Li and Fong 2019)	Coefficient of Variation for Diabetics Prediction	Pima diabetes data, LR, DT, RF, SVM, NB, KNN	Accuracy, Sensitivity, Specificity, F1 Score, AUC	Achieved an accuracy of 0.85, sensitivity of 0.80, specificity of 0.90, F1 score of 0.82, AUC of 0.89	The generalizability of the proposed method to other disease or medical conditions not discussed.

Working Methodology

System Architecture

To explore critical factor analysis in detail, we have introduced a methodology called Inclusive Feature Selection Strategy (IFSS) in order to investigate and analyze the diabetes mellitus data, as represented in Figure 2. The constructed architecture for estimating the critical factors of diabetes mellitus with standardized and non-standardized processes using feature selection methods is shown below. The working behavior of the proposed system architecture consists of following major components such as: University of California Irvine (UCI) repository data set, cleaning and reduction components, balancing component, standardized and non-standardized modules, model analysis and evaluation on every obtained set individually, feature

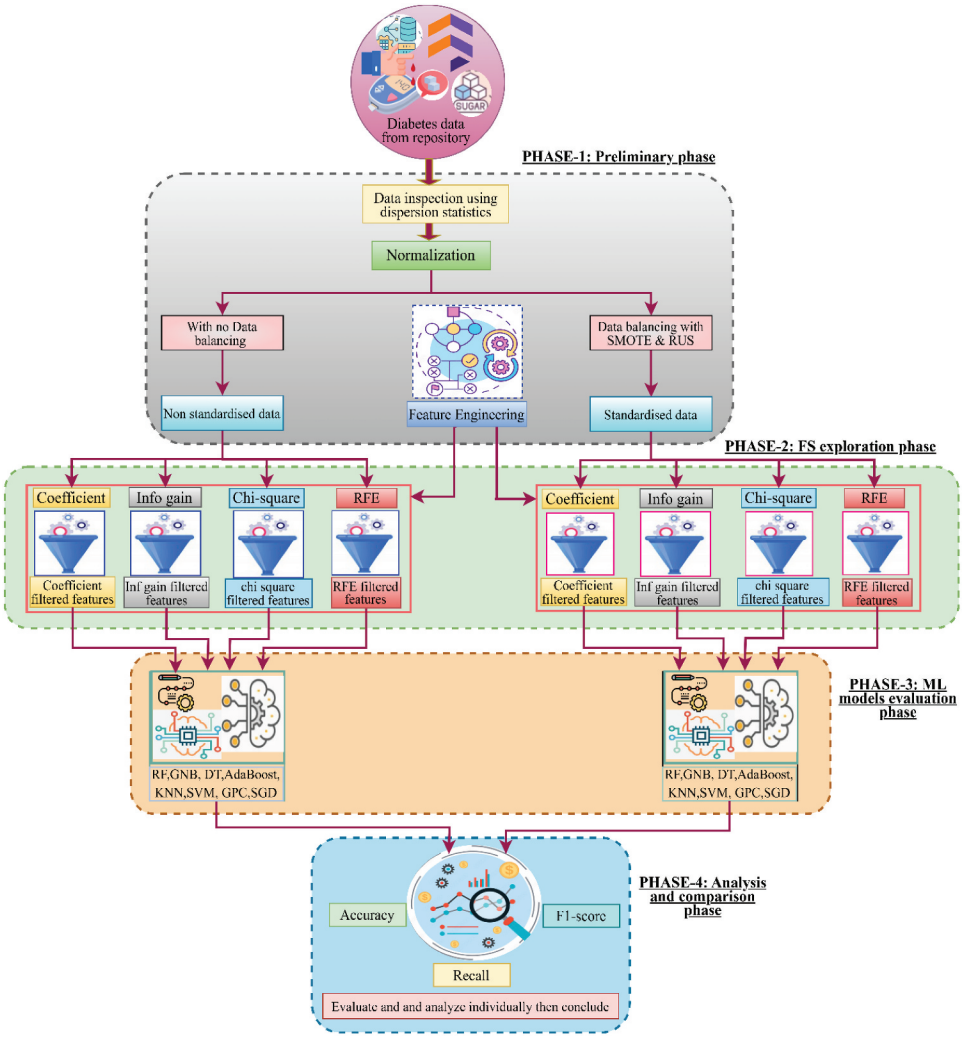

Figure 2. The proposed inclusive feature selection strategy (IFSS) working methodology.

selection methods comparison based on classification results. The proposed system has been organized into 4 phases.

The phase-1 represents the preliminary phase involves data collection, preparation and determining basic characteristics of data using data dispersion concepts. Phase-2 deals with feature selection exploration phase. Similarly, phase 3 and phase-4 organized as model analysis phase, comparison phase respectively. The activities such as data collection, cleaning, reduction, and balancing the label data for analysis were implemented in Phase 1. Initially the phase-1 comprises of the key operations, such as dealing with missed data, applying normalization, and balancing, etc. The concept of preprocessing would help us to get the coherent data to run models in a sophisticated manner without complexity. Obviously more imperative than good models preprocessing should be treated as a paramount important concept to deal with, unstructured, ambiguous, missing or error filled data. The basic statistical characteristics of the data has been explored using dispersion methods Inter Quartile Range (IQR), standard deviation, skewness and kurtosis etc. The results of the dispersion statistics are shown in Figure 3. The data doesn't contain missing data and in the identified null positions we didn't replace the values with data missing handling techniques such as mean, median, the nearest value, and the most repeated value. Because by filling the missed data with the above-mentioned techniques could lead to manipulation of actual data and causes to get falsified/biased results. Phase 2 has been divided into two modules such as standardized and non-standardized. The feature selection methods (RFE, Info gain, correlation, Chi-square) are applied to both modules but the standardized module contains the balanced data whereas non-standardized module experimented with no balancing. The sampling method (Bach, Werner, and Palt 2019; Zhu et al. 2020) was used to balance the dataset. In phase 3, the KNN (Ali, Neagu, and Trundle 2019; Guo et al. 2003; Mucherino, Papajorgji, and Pardalos 2009) model was applied and analyzed individually over obtained sets from the phase-2 standardized and the non-standardized modules. Later in phase 4

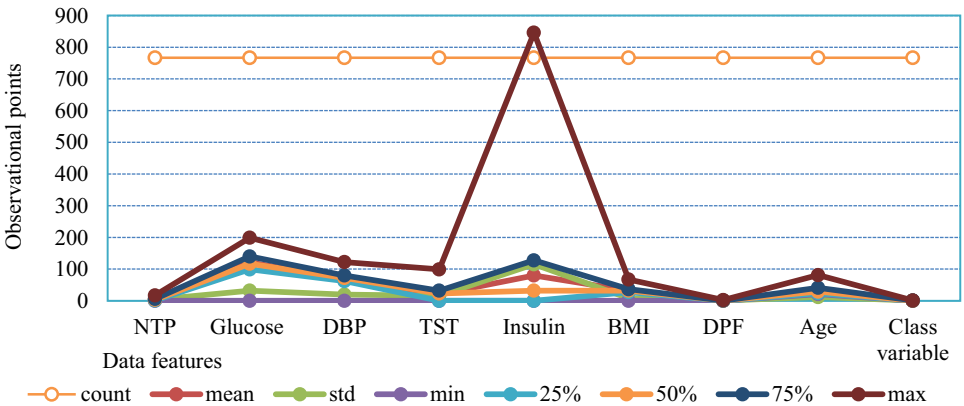

Figure 3. Descriptive statistics of diabetes.

the classification results were computed by considering the metrics accuracy, precision, recall, F1score.

Implemented FS Methods

To understand the concept of consistency in the feature selection methods, we worked with four methods by using the dataset found in the NIDDK data repository. The applied methods are Information Gain (IG) (Di Franco 2019), chi-square (Baker and Cousins 1984; Ottenbacher 1995), RFE (Bahl et al. 2019; Huerta et al. 2013), and correlation (Freund, Wilson, and Mohr 2010) strategies. To show the consistency of feature selection methods and to explore diabetes critical factors characteristics the methodology was implemented by employing the feature selection methods and ML techniques over the non-standardized and standardized data.

Information Gain

It is necessary to know the amount of information contained in the features in the form of classes. Information gain helps to define the possibility of the occurrence of a class or surprise with respect to the target variables. For performing feature selection, IG (Di Franco 2019) helps by evaluating each variable based on variable gain. Mutual information estimates the gain value with the help of two variables.

The Information gain can be computed based on below given formula.

$$Information\,gain = Entrophy\,(parent) - Average\,(Entrophy\,(child)) \quad (1)$$

Information Gain will calculate the entropy reduction by dividing a dataset with respect to the considered value of a randomly selected variable. Entropy indicates the uncertainty or impurity that exists in the data set. To understand clearly about IG behavior the Table 3 represented below.

$$Entrophy = \sum_{i=1}^{n} -probability(class_i)\log_2((probability(class_i))) \quad (2)$$

Whereas n represents the number of different class values

Chi-Square

Chi-Square (Baker and Cousins 1984; Ottenbacher 1995) is a simple method for performing feature selection to implement the classification task. Feature selection aims to focus on selecting the highly dependent variables. The Chi-square method is used to test whether two variables are related to each other or independent of one

Table 3. Representation of information gain behavior based on value.

Entropy	Info gain	Surprise possibility
Lower	High	More
Higher	Low	Less

another. In simple terms, the chi-square method determines the relationship between dependent and independent variables.

If the two features are dependent and the observed values are not close to the expected values, then the chi-square value is high. If the two features are independent, then the observed values are close to the expected values, then the chi-square value is low. A higher score means both are dependent, one can select for model training. A lower chi-square means that both variables are independent of each other and cannot be selected for experimentation. To interpret chisquare easily we have represented in Table 4.

$$\chi_2 = \sum_{i=1}^{n} \frac{(OV_i - EV_i)^2}{EV_i} \tag{3}$$

where *OV and EV* represent observed and expected values and n indicates the number of instances.

Pearson Correlation Coefficient

The concept of correlation was introduced by Francis Galton and developed by Pearson Karl to establish the relationship between variables of data sets. The correlation values always lie between − 1 and + 1 including zero. The meaning of the relationship between variables is defined based on the obtained coefficient values. The meaning and relationships among variables based on obtained coefficient values are represented in Table 5 given below:

The Pearson correlation (Pearson's Correlation Coefficient 2008; Ratner 2009) calculates the strength of the relation between two features. It ranges between −1 and 1. The value of − 1 means a complete negative correlation, 0 indicates no correlation, whereas + 1 means a total positive correlation

Table 4. Representation of chi square behavior based on value.

Chi-Square value	Dependent and Independent variable relationship	Select model training
Smaller value	Observed feature not dependent on the response feature	Do not select (No)
Higher value	Observed feature more dependent on the response feature	Select (Yes)

Table 5. Representation of coefficient strength features.

Coefficient Value	Meaning	Features consider
−1	Total Negative Perfect Correlation	No
−0.9 To −0.7	Strong Negative Correlation	Yes
−0.6 To −0.4	Partial Negative Correlation	Yes
−0.3 and −0.1	Lower Negative Correlation	Yes
0	No Correlation	No
0.1 To 0.3	Partial Positive Correlation	Yes
0.4 To 0.6	Lower Positive Correlation	Yes
0.7 To 0.9	Strong Positive Correlation	Yes
+1	Total Positive Perfect Correlation	No

$$Reg\_coef(\rho) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{4}$$

Exhaustive Feature Selection Based RFE

The RFE approach comes under a wrapper mechanism where there is a separate learner for selecting the features. In general, the selection of the learner depends upon the developer's choice. RFE (Bahl et al. 2019; Huerta et al. 2013) works based on the brute force evaluation concept. It implies that it will try on all possible feature combinations and then produces the best subset. However, this method is expensive and needs more time compared to forward selection and backward elimination. The reason for computational complexity in RFE is because that it considers every possible combination of features.

Description of the Data Set Used for Analysis

The data set contains diagnostic-related measurements and other pertinent information shown in Table 6, which help us to predict whether the person has the disease or not. The data set used is Pima (Kaggle 2022), which contains 768 records and 9 features. The target feature for classification has two classes either diabetic or not. The Indian Pima dataset is developed by the NIDDK, Vincent Sigillito from Johns Hopkins University. The features are plasma glucose concentration (PGC), BMI (Body Mass Index), DPF (Diabetic Pedigree Function), Diastolic Blood Pressure (DBP), age, 2-hour serum insulin, skin thickness, Triceps Skinfold Thickness (TSFT), Number of Times Pregnant (NTP) and class variable.

Coding Design Explanation

To Estimate the critical factors of diabetes and to observe the consistency of feature selection methods the coding part is carried out in two phases. Phase 1 deals with unbalanced data by applying four feature selection methods. In phase 2, the standardized data has been analyzed by applying feature selection methods. Later, the ML model was applied to both phases of selected features. The performance of feature selection methods is examined based on the obtained results after applying the model.

Table 6. Data set description.

Data Characteristics	Associated information
Data set Name/Title	Pima Indians Diabetes Database
Data set type/Data type	Multivariate
Sources and data owned by	NIDDK, The Johns Hopkins University
Labelled feature classes	1 and 0 are interpreted as tested +ve and -ve for diabetes

Phase-1 Implemented algorithm

Input: Dependent and Independent variables (*DV and IV's) of* Pima unbalanced data set.

Output: computed the coefficients values with selected features and generated data frames df2, df3, df4, df5.

Algorithm 4.1: **Algorithm for Non-Standardized Data**

```
1. Begin
2. Include the libraries such as seaborn, pandas, and NumPy.
3. Select the chi^2, SelectKBest, mutual info, and RFE from sklearn
4. Read the file using "pandas. read" format
5. Divide the data set into x and y using the syntax
      x=dataframe. drop (['label feature'], axis=1), y=dataframe ['label feature']
6. Apply feature selection methods subsequently then generate frames
7. def coefficients(IV, DV, k):
8. # Calculate mutual information between IV, DV
9. mi = mut_inf(IV, DV)
10. return mi
11. construct the df4 frame using assign () method based on the selected features
12. #Calculate Pearson correlation between IV's and DV
13. corr = np.array(pearson(IV, DV))corr = []
      for i in range(IV.shape[]):
      corr.append(IV, DV)
      end for
      corr = np.array(corr)
14. return corr
15. construct df2 set frame using assign() method based on selected features
16. #Calculate chi-square between IV's and DV
17. chi2_scores= chi2(IV, DV)
18. return chi2_scores
19. construct df3 dataset frame using assign() method based on selected list
20. #Use recursive feature elimination to select the top k features
21. selector = RFE(SelectKBest(k=k), features=k)
22. selector.fit(IV, DV)
23. return selector.estimator.coef.
24. construct df5 data frame using assign() method based on selected features
25. end def.
26. List all feature selection based generated data frames.
```

Results Discussion

Estimation of the critical factors for diabetes mellitus with the standardized and non-standardized process is implemented with info gain, chi-square, RFE, and coefficient. Significant frames are constituted based on feature selection methods before and after standardization based on features sets shown in Table 7 The following methods are employed to the generated frames to conduct our experimentation: AdaBoost, RF, DT, GPC, SVM, and SGD respectively. The evaluation metrics such as accuracy, recall, and F1 score are used for evaluating the implemented models.

The above shown Figure 3 quickly enables us to comprehend the general characteristics of diabetes data. It presents a summary of the dispersion of diabetes features, including measures such as central tendency, mean, standard deviation, and IQR.

Table 7. Representation of features considered for data frames construction for phase 1 and phase 2.

FS Method	Filtered Features		Prior and Post status in selection of variables
	Before Standardizing Data	After Standardizing Data	
Info gain	NTP, PGC, Insulin, BMI, Age	NTP, PGC, TSFT, Insulin, BMI	Yes
Chi square	DBP, DPF, NTP, BMI, TSFT	NTP, BMI, TSFT, DPF	Yes
RFE	NTP, PGC, Age, BMI, DPF	NTP, PGC, Age, BMI, DPF	No
Coefficient	Insulin, Age, BMI, DPF, TSFT	Insulin, Age, BMI	Yes

Table 8. Feature selection methods computed coefficient scores.

Mutual information		Correlation		Chi-square		RFE Ranking		
Feature	Value	Feature	Value	Feature	Value	Feature	Support	Rank
Glucose	0.139614	NPT	0.221898153	TST	3.60E−14	NTP	True	1
Age	0.06401	Glucose	0.466581398	Glucose	1.62E−221	Glucose	True	1
BMI	0.072143	DBP	0.06506836	NTP	1.01E−15	DBP	False	2
DPF	0.020957	TST	0.074752232	DBP	1.01E−15	TST	False	3
NTP	0.036297	Insulin	0.130547955	DPF	3.67E−02	Insulin	False	4
TST	0.018813	BMI	0.292694663	BMI	4.06E−22	BMI	True	1
Insulin	0.067242	DPF	0.173844066	Age	7.64E−29	DPF	True	1
DBP	0.005889	Age	0.238355983	Insulin	0.00E+00	Age	True	1

Table 8 presents the calculated coefficients scores for MI, chi square, RFE, and coefficient methods. The computed coefficient values will assist us in selecting the optimal features for conducting analysis.

FS Methods Based Obtained Models Results of with Respect to Prior and Post Standardization

Based on the results shown in Table 9, we have created Figure 4 to visually represent our findings. The Figure 4a–d, e–h illustrate the results of the ML models before and after standardization, specifically

Algorithm 4.2: Algorithm for ML models then apply to obtained results of 4.1 algorithm

```
(1)  Import the necessary package libraries such as seaborn, pandas, scikit
(2)  Include all the frames cyclically for data frames df2, df3, df4, df5 subsequently by applying the model
(3)  Use sklearn import train_test_split and metrics to implement ML model on frames
(4)  While frame df2 ≤ df5 do
(5)     //df2, df3, df4, df5 the associated sets are x2-y2, x3-y3, x4-y4, x5-y5 sets
(6)          for each corresponding frames Split into training and testing sets
(7)          x_trainset, x_testset, y_trainset, y_testset = trainset_testset_split(x2, y2)
(8)          Import and train the model on the training set by specifying k value
(9)               ML_model = classifier name (),
(10)              ML_model.fit (x_trainset, y_trainset)
(11)         Do predictions using the testing set
(12)              y_pred = ML_model.predict (x_test)
(13)         Compare the actual versus predicted values
(14)    df = df +1;
(15) end while
(16) return the results of all models obtained on all frames
(17) Display all the accuracy, recall, and F1 scores for all the cases.
```

Phase-2 implemented algorithm

Input source	Pima balanced data set
Output	Selected features and generated data set frames df7, df8, df9, df10

Algorithm 4.3: Algorithm for Standardized Data

```
(1) Include the required packages and read the file
(2) Divide the data set into x and y using the syntax x=dataframe. drop(['label feature'], axis=1), y=dataframe
    ['label feature']
(3) Balance data set using the sampling process
(4) Import RandomUnderSampler from imblearn and apply to data frames x and y
(5)         RUS=RandomUnderSampler (sampling strategy=1)
(6)         x_res,y_res=RUS.fit_resample(x, y)
(7) Check label feature classes are balanced or not using "feature name.value_counts()
(8) Replace with the missing values in df6 data frame
(9)         df6=df6.fillna (df6.mode ().iloc [0])
(10) Check whether the data is standardized properly or not by looking overall data set
(11) Apply feature selection methods subsequently then generate frames
(12)        Correlation method
(13)                correlation(dataset, threshold):
(14)                col_corr = set()
(15)                corr_matrix = dataset.corr()
(16)                return col_corr
(17)                select the corr_features
(18)                construct df7 set frame using assign() method based on selected features
(19)        Chi square
(20)                f_p_values=chi2(x,y)
(21)                p_values=pd. series(f_p_values[1])
(22)                p_values. Index=columns
(23)                Listing features scores P_values
(24)                construct df8 dataset frame using assign() method based on selected list
(25)        mutual information
(26)                mutual info = mutual_info_classif(x, y)
(27)                mutual info based select the top 5 important features
(28)                sel_five_cols = SelectKBest(mutual_info_classif, k=5)
(29)                sel_five_cols.fit (x, y)
(30)                construct a df9 frame using assign() method based on selected features
(31)        RFE using the SVR estimator
(32)                Import SVR method from sklearn.SVM
(33)                consider SVR as an estimator and RFE as selector i.e., estimator = SVR()
(34)                selector = RFE (estimator, top_features=5)
(35)                Fit x,y to the selector as selector.fit (x, y)
(36)                Listing feature scores and selection using ranking and selector
(37)                construct df10 data frame using assign() method based on the selected features
(38)    Return the results obtained on all frames
(39) Use the same code part of the 4.2 algorithm on these generated frames and apply the ML model
(40) Display the accuracy of the model, and then visualize the FS based results
```

regarding Mutual Information (MI), Correlation, Chi-square, and Recursive Feature Elimination (RFE). Furthermore, it displays the distinct feature selection methods of implemented machine learning models for the results of 4.a and 4.e, with regards to mutual information. It also includes corresponding results for correlation (4.b and 4.f), chi square (4.c and 4.g), and RFE (Figure 4d–h) with all combinations before and after standardization.

Table 9. The obtained classification results before and after standardization.

Before standardisation				After standardisation			
Model name	Accuracy	Recall	F1-score	Model name	Accuracy	Recall	F1-score
Mutual information				**Mutual information**			
RF	0.747664	0.830721	0.732469	RF	0.747664	0.830721	0.732469
GNB	0.747664	0.830721	0.732469	GNB	0.747664	0.830721	0.732469
DT	0.745327	0.5	0.427041	DT	0.745327	0.5	0.427041
AdaBoost	0.745327	0.5	0.427041	AdaBoost	0.745327	0.5	0.427041
KNN	0.745327	0.5	0.427041	KNN	0.745327	0.829154	0.730309
SVM	0.745327	0.5	0.427041	SVM	0.745327	0.5	0.427041
GPC	0.745327	0.5	0.427041	GPC	0.745327	0.5	0.427041
SGD	0.735981	0.49373	0.423957	SGD	0.712617	0.80721	0.700231
Correlation				**Correlation**			
AdaBoost	0.739414	0.698618	0.693337	RF	0.747664	0.830721	0.732469
RF	0.70684	0.646313	0.646313	GNB	0.747664	0.830721	0.732469
GNB	0.70684	0.607296	0.613285	DT	0.745327	0.5	0.427041
SVM	0.703583	0.569227	0.567973	AdaBoost	0.745327	0.5	0.427041
KNN	0.687296	0.638991	0.634161	SVM	0.745327	0.5	0.427041
GPC	0.674267	0.629775	0.62322	GPC	0.745327	0.5	0.427041
Decision Tree	0.648208	0.588582	0.586005	KNN	0.742991	0.498433	0.426273
SGD	0.560261	0.542627	0.526478	SGD	0.721963	0.81348	0.708799
Chi-square				**Chi-square**			
GNB	0.723127	0.654583	0.657963	RF	0.747664	0.830721	0.732469
SGD	0.716612	0.532924	0.492543	GNB	0.747664	0.830721	0.732469
SVM	0.690554	0.507988	0.462386	KNN	0.747664	0.830721	0.732469
AdaBoost	0.684039	0.649693	0.639414	DT	0.745327	0.5	0.427041
RF	0.654723	0.59319	0.591222	AdaBoost	0.745327	0.5	0.427041
KNN	0.641694	0.564465	0.564864	SVM	0.745327	0.5	0.427041
GPC	0.631922	0.57381	0.570619	GPC	0.745327	0.5	0.427041
DT	0.62215	0.579903	0.572095	SGD	0.740654	0.496865	0.425503
RFE				**RFE**			
AdaBoost	0.778502	0.73277	0.723277	RF	0.747664	0.830721	0.732469
GNB	0.76873	0.716103	0.718199	GNB	0.747664	0.830721	0.732469
RF	0.765472	0.713799	0.715184	Decision Tree	0.745327	0.5	0.427041
KNN	0.758958	0.692934	0.69902	AdaBoost	0.745327	0.5	0.427041
SVM	0.758958	0.663671	0.676492	KNN	0.745327	0.5	0.427041
DT	0.700326	0.654711	0.649404	SVM	0.745327	0.5	0.427041
GPC	0.700326	0.622197	0.625808	GPC	0.745327	0.5	0.427041
SGD	0.306189	0.509217	0.247107	SGD	0.735981	0.49373	0.423957

Mutual Information

Based on graphs 4.a and 4.e, as well as Table 9, we can observe that most of the models achieve an accuracy of 74%, with the exception of SGD. Surprisingly, prior to standardization, the SGD model outperformed with an accuracy of 73%, compared to 71% after standardization. Nonetheless, SGD demonstrated strong recall with F1 score scores of 80% and 70% after standardization. Among all listed models, the SGD model has the lowest performance with accuracy rates of 0.73 and 0.71 before and after standardization. Only two models RF and GNB outperformed in terms of recall and F1 score, but after standardization, four models RF, GNB, KNN, and SGD showed improved results of 83%, 83%, 82%, and 80%.

Figure 4. The figures from a–d, e–h represents different ML models obtained results with respect to MI, correlation, Chi square, RFE before standardization, after standardization.

Correlation Results

Based on Figure 4b–f and Table 9, both the RF and GNB models have the highest accuracy of 0.74, recall of 0.83, and F1 score of 0.73 after standardization. The KNN model has the highest recall of 0.63 and F1 score of 0.63 before standardization, but after standardization, the recall drops to 0.49 and the F1 score drops to 0.42. The SGD model has the lowest performance among all listed models, with accuracy, recall, and F1 score rates of 0.72, 0.81, and 0.70, respectively, after standardization. The standardization process resulted in all models achieving an accuracy of 74%, with the exception of the SGD model.

Chi Square

Figures 4c–g and Table 9 of the chi square-based results revealed that all the models showed better performance with an accuracy of above 0.74. Significantly, the RF, GNB, and KNN models obtained good recall scores (0.83) and F1 scores (0.73) after standardization. The DT model has the lowest performance among all listed models, with an accuracy rate of 0.62, a recall rate of 0.57, and an F1 score of 0.57 before standardization.

RFE Results

Based on RFE results in Figure 4d–h, and Table 9, indifferently few models such as Adaboost, GNB, RF, KNN, and SVM performed better before standardization in terms of accuracy only. The RF and GNB models yield the highest recall of 0.83 and the F1 score of 0.73 after standardization. The SGD model has the lowest performance among all listed models, with accuracy, recall, and F1 scores before and after standardization.

The major characteristics of the models we have identified are as follows on an overall basis.

(i) The RF, GNB, and AdaBoost models have demonstrated strong performance in terms of prior and post standardization.
(ii) The correlation, MI, and Chi square feature selection factors have generally shown significant improvement after standardization in the obtained ML results. However, the RFE-based model results did not demonstrate the same level of improvement.
(iii) However, both the SGD and GPC models yielded unsatisfactory results in both scenarios, i.e., prior and post standardization.
(iv) The equal accuracy rate was obtained for all models with respect to all MI, correlation, RFE, and chi square after standardization.

SMOTE Based Obtained Results

The performance parameters used to assess the accuracy, recall, and F1 score of different machine learning models are displayed in Figure 5. The RF model has the greatest accuracy (0.83), recall (0.82), and F1 score (0.81). With an accuracy of 0.70, recall of 0.70, and F1 score of 0.70, the DT model performed moderately. The GPC model, with an accuracy of 0.78, recall of 0.78, and F1 score of 0.77, outperforms DT but falls short of RF. The AdaBoost model has an accuracy rate of 0.76, a recall rate of 0.76, and F1 score of 0.76. It performs better than both DT and SVM, but not as well as RF or GPC. The GNB model's metrics are close to those for AdaBoost, with an accuracy rate of 0.73, a recall rate of 0.72, and F1 score of 0.72. The accuracy rate of 0.76, recall rate of 0.76, and F1 score of 0.75 of the KNN model are comparable to those of AdaBoost. Finally, the SGD model has the lowest performance among all listed models, with an accuracy rate only reaching 0.53.

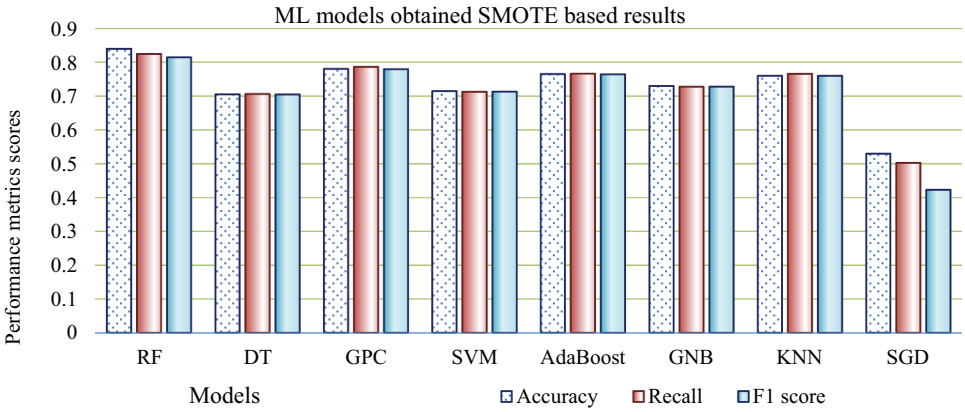

Figure 5. The obtained SMOTE based classification results with standardization.

Comparison Results

Table 10 shows the accuracy results of various machine learning models applied by different authors. From the represented table, it can be seen that the proposed concept model has the highest accuracy rate of 83% among all the models listed. The Saxena R et al. (Saxena et al. 2022) strategy uses KNN, RF, DT, and Multilayer Perceptron (MLP) algorithms, and RF has a better accuracy rate of 79.80%. The Sabitha E et al. approach (Sabitha and Durgadevi 2022) uses LR, RF, DT, SVC, GNB, and KNN algorithms, and LR has shown an improved accuracy rate of 80%. The Kumari S et al. methodology (Kumari, Kumar, and Mittal 2021) uses a soft voting classifier (SVC), LR, NB, RF, XGB, NB, and DT algorithms, and SVC obtained good accuracy at 79%. The Joshi R et al. (Joshi and Dhakal 2021) employs the LR algorithm and has an accuracy rate of 78.26%. Finally, Chatrati S et al. (Chatrati et al. 2022) methodology obtained an accuracy rate of 72% by leveraging the LDA algorithm.

The proposed IFSS methodology has evolved through various activities such as sampling, assessing statistical characteristics, utilizing FS methods, ML models, and evaluating metrics to perform critical factor analysis for diabetes mellitus. The critical factor analysis revealed that the determined primary influential factors are NTP, PGC, BMI, and age. In addition, the recognized secondary features are TSFT, insulin, and DPF. Further, DBP was identified as the least influential feature overall.

Table 10. The comparative results with other state-of-the-art methods.

Authors	Models applied	Accuracy
Saxena R et al. (Saxena et al. 2022)	KNN, **RF**, DT, MLP	79.80%
Chatrati S et al. (Chatrati et al. 2022)	**LDA**	72%
Sabitha E et al. (Sabitha and Durgadevi 2022)	**LR**, RF, DT, SVC, GNB, KNN	80%
Joshi R et al. (Joshi and Dhakal 2021)	**LR**	78.26%
Kumari S et al. (Kumari, Kumar, and Mittal 2021)	**Soft voting classifier**, LR, NB, RF, XGB, NB, DT	79%
Proposed concept	**RF**, DT, GPC, SVM, AdaBoost, **GNB**, KNN, SGD	83%

Time Complexity Analysis

The PCC algorithm time complexity is $O(n)$, where n is the number of elements in the input. The time complexity of the chi-square and mutual information algorithms is $O(n)^2$. The time complexity of the RFE algorithm is $O(k*(n)^2$, where k is the number of features to select and n is the number of samples. To understand the complexity analysis easily we have depicted Figure 6.

The overall complexity would be

$$\begin{aligned} &= \text{correlation complexity} + \textit{MI complexity} + \textit{Chi square complexity} \\ &\quad + \textit{RFE complexity} \\ &= O(n) + O(n)^2 + O(n)^2 + O\left(k*(n)^2\right) = O\left(k*(n)^2\right) \end{aligned}$$

Therefore, the overall time complexity to calculate coefficients by different methods would be $O(k(n)^2)$, where k is the number of features to select and n is the number of samples.

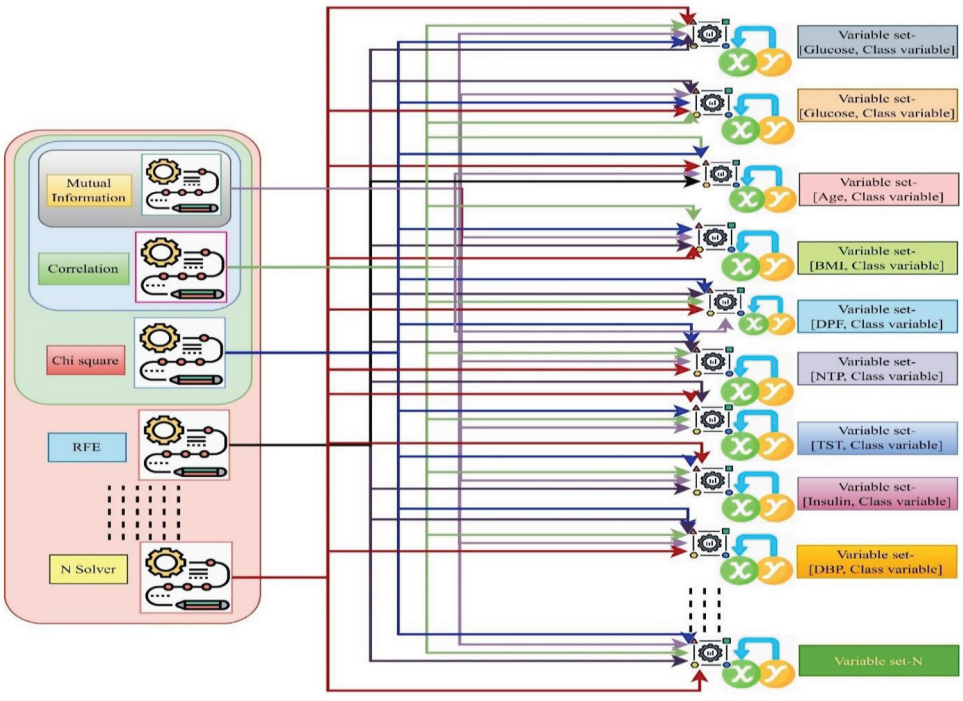

Figure 6. The computational process of feature selection methods with dependent variable and independent variables for coefficients determination.

Conclusion

Despite substantial progress in the treatment and prevention of diabetes mellitus, the disease continues to be a major public health concern worldwide. The article focused on experimenting with various mechanisms, including sampling, feature selection exploration, applying different ML models, and evaluating metrics on the Pima data before and after standardization. In this research, we applied a number of feature selection techniques to a comprehensive examination of estimated essential components in diabetes mellitus utilizing both conventional and nonstandard mechanisms. A small crucial set of features in the dataset facilitates easy interpretation of the model, yields better prediction, optimizes training time and computational cost, as well as mitigates the risk of overfitting. To isolate critical elements linked to diabetes mellitus, we used Info gain, chi-square, Recursive Feature Elimination (RFE), and coefficient. The RF, GNB, and AdaBoost models have demonstrated strong performance in terms of prior and post standardization. Info gain, chi-square, and coefficient all improved greatly in their ability to pinpoint influential elements after being subjected to standardization. Nonetheless, interestingly, the RFE approach showed resistance to standardization, with performance being constant both before and after the process. However, the RFE-based model results did not demonstrate the same level of improvement. However, both the SGD and GPC models yielded unsatisfactory results in both scenarios, i.e., prior and post standardization. These results highlight the need of using suitable feature selection methods and considering standardization in order to enhance the precision of crucial factor identification in diabetes mellitus. Better public health outcomes may result from the study's findings helping researchers and healthcare providers create more effective techniques for managing and preventing diabetes mellitus.

Future Work

This work sheds light on several promising new directions to investigate in the mysterious world of diabetes mellitus research. The following directions show promise for further exploration to continue to delve into the depths of this mysterious realm. In order to uncover even further improvements in performance across all feature selection methods, future work could investigate on alternate data transformation strategies or fine-tune existing standardization procedures. The identification of crucial components inside the intricate web of diabetes mellitus may require future research into Recursive Feature Addition (RFA), or evolutionary algorithms. Additionally, validation and clinical relevance is crucial to verify the findings in clinical settings as the miraculous discoveries develop. The clinical relevance and application of the identified essential criteria may be assessed in future study through collaboration with

healthcare practitioners and specialists. This verification will help those fighting diabetes mellitus by closing the gap between theoretical considerations and practical applications. It will yield better results in using ML in diabetes research subsequently, work can be carried out with the real-time data set to extrapolate and develop an automatic recommendation system. This can help the diabetic affected protagonists to decide on their diet.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

To carry out research work the funding has been granted by the Vellore Institute of Technology in Vellore, Tamil Nadu, India.

ORCID

L. D. Dhinesh Babu http://orcid.org/0000-0002-3354-8713

Author Contributions

The first author carried out the methodology, data curation, experiments, and draft preparation. In addition to providing supervision, the coauthor also helped with conceptualization, validation, review, and editing.

Availability of Data And Materials

Upon a reasonable request, the sources and other pertinent information will be provided.

Consent for Publication

The manuscript not contains any individual person's data in any form.

Ethics Approval And Consent To Participate

This manuscript does not contain any studies with human participants or animals performed by the author.

References

10 Surprising Things That Can Spike Your Blood Sugar | CDC. Accessed May 24, 2023. [Online]. Available: https://www.cdc.gov/diabetes/library/spotlights/blood-sugar.html

7th edition | IDF Diabetes Atlas. Accessed Dec 11, 2023. [Online]. Available: https://diabetesatlas.org/atlas/seventh-edition/

Aha, D. W., D. Kibler, M. K. Albert, and J. R. Quinian. Jan, 1991. Instance-based learning algorithms. *Machine Learning* 6 (1):37–66. doi:10.1007/BF00153759.

Alalwan, S. A. D. Apr, 2019. Diabetic analytics: Proposed conceptual data mining approaches in type 2 diabetes dataset. *Indonesian Journal of Electrical Engineering and Computer Science* 14 (1):88–95. doi:10.11591/IJEECS.V14.I1.PP88-95.

Alam, S., M. K. Hasan, S. Neaz, N. Hussain, M. F. Hossain, and T. Rahman. Apr, 2021. Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management. *Diabetology 2021* 2(2):36–50. doi:10.3390/DIABETOLOGY2020004.

Ali, N., D. Neagu, and P. Trundle. Dec, 2019. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences* 1(12):1–15. doi:10.1007/s42452-019-1356-9.

Alyasiri, O. M., Y. N. Cheah, A. K. Abasi, and O. M. Al-Janabi. 2022. Wrapper and hybrid feature selection methods using metaheuristic algorithms for English Text Classification: A systematic review. *IEEE Access* 10:39833–52. doi:10.1109/ACCESS.2022.3165814.

Ang, J. C., A. Mirzal, H. Haron, and H. N. A. Hamed. Sep, 2016. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology & Bioinformatics / IEEE, ACM* 13 (5):971–89. doi:10.1109/TCBB.2015.2478454.

Bach, M., A. Werner, and M. Palt. 2019. The proposal of undersampling method for learning from imbalanced datasets. *Procedia Computer Science* 159 (Jan):125–34. doi:10.1016/J.PROCS.2019.09.167.

Bahl, A., Hellack, B., Balas, M., Dinischiotu, A., Wiemann, M., Brinkmann, J., Luch, A., Renard, B. Y., Haase, A. Mar, 2019. Recursive feature elimination in random forest classification supports nanomaterial grouping. *NanoImpact* 15:100179. doi: 10.1016/J.IMPACT.2019.100179.

Baker, S., and R. D. Cousins. Apr 1984. Clarification of the use of CHI-square and likelihood functions in fits to histograms. *Nuclear Instruments and Methods in Physics Research* 221 (2):437–42. doi:10.1016/0167-5087(84)90016-4.

Bommert, A., X. Sun, B. Bischl, J. Rahnenführer, and M. Lang. 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis* 143 (Mar):106839. doi:10.1016/J.CSDA.2019.106839.

Browne, M. W. Mar 2000. Cross-validation methods. *Journal of Mathematical Psychology* 44 (1):108–32. doi:10.1006/JMPS.1999.1279.

Buyrukoğlu, S., and A. Akbaş. Apr 2022. Machine learning based early prediction of type 2 diabetes: A new hybrid feature selection approach using correlation matrix with heatmap and SFS. *Balkan Journal of Electrical and Computer Engineering* 10 (2):110–17. doi:10.17694/BAJECE.973129.

Chang, V., J. Bailey, Q. A. Xu, and Z. Sun. Aug, 2023. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing & Applications* 35(22):16157–73. doi: 10.1007/s00521-022-07049-z.

Chart: Where Diabetes Burdens Are Rising | Statista. Accessed May 24, 2023. [Online]. Available: https://www.statista.com/chart/23491/share-of-adults-with-diabetes-world-region/

Chatrati, S. P., G. Hossain, A. Goyal, A. Bhan, S. Bhattacharya, D. Gaurav, and S. M. Tiwari. Mar, 2022. Smart home health monitoring system for predicting type 2 diabetes and hypertension. *Journal of King Saud University - Computer and Information Sciences* 34 (3):862–70. doi:10.1016/J.JKSUCI.2020.01.010.

Chen, R. C., C. Dewi, S. W. Huang, and R. E. Caraka. Dec, 2020. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data* 7(1):1–26. doi:10.1186/s40537-020-00327-4.

Choubey, D. K., P. Kumar, S. Tripathi, and S. Kumar. Dec, 2020. Performance evaluation of classification methods with PCA and PSO for diabetes. *Network Modeling Analysis in Health Informatics and Bioinformatics* 9(1):1–30. doi:10.1007/s13721-019-0210-8.

Chowdary, P. B. K., and R. U. Kumar. 2021. Diabetes Classification using an Expert Neuro-fuzzy Feature Extraction Model. *International Journal of Advanced Computer Science and Applications* 12 (8):368–74. doi:10.14569/IJACSA.2021.0120842.

Dalianis, H. 2018. Evaluation Metrics and Evaluation. *Clinical Text Mining* 45–53. doi:10.1007/978-3-319-78503-5_6.

De Silva, K., D. Jönsson, and R. T. Demmer. Mar, 2020. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. *Journal of the American Medical Informatics Association: JAMIA* 27 (3):396–406. doi:10.1093/JAMIA/OCZ204.

Diabetes Prevalence Expected to Double Globally by 2050. 2023. Accessed Dec 13, 2023. [Online]. Available: https://www.ajmc.com/view/diabetes-prevalence-expected-to-double-globally-by-2050

Di Franco, A. 2019. Information-gain computation in the fifth system. *International Journal of Approximate Reasoning* 105 (Feb):386–95. doi:10.1016/J.IJAR.2018.11.013.

Doğru, A., S. Buyrukoğlu, and M. Arı. Mar, 2023. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. *Medical & Biological Engineering & Computing* 61 (3):785–97. doi:10.1007/s11517-022-02749-z.

Fei, Z., F. Yang, K. L. Tsui, L. Li, and Z. Zhang. 2021. Early prediction of battery lifetime via a machine learning based framework. *Energy* 225 (Jun):120205. doi:10.1016/J.ENERGY.2021.120205.

Freund, R. J., W. J. Wilson, and D. L. Mohr. 2010. Nonparametric methods. *Statistical Methods* 689–719. doi:10.1016/B978-0-12-374970-3.00014-7.

Gromova, L. V., S. O. Fetissov, and A. A. Gruzdkov. Jul, 2021. Mechanisms of glucose absorption in the small intestine in health and metabolic diseases and their role in appetite regulation. *Nutrients* 13(7). doi:10.3390/NU13072474.

Guo, G., H. Wang, D. Bell, Y. Bi, and K. Greer. 2003. KNN model-based approach in classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2888:986–96. doi:10.1007/978-3-540-39964-3_62/COVER/.

Gupta, S. C., and N. Goel. 2023. Predictive modeling and analytics for diabetes using hyper-parameter tuned machine learning techniques. *Procedia Computer Science* 218:1257–69. doi:10.1016/J.PROCS.2023.01.104.

Gupta, A., I. S. Rajput, Gunjan, V. Jain, and S. Chaurasia. Sep, 2022. NSGA-II-XGB: Meta-heuristic feature selection with XGBoost framework for diabetes prediction. *Concurrency & Computation: Practice & Experience* 34 (21):e7123. doi:10.1002/CPE.7123.

Gürsoy, M. İ., and A. Alkan. Dec, 2022. Investigation of diabetes data with permutation feature importance based deep learning methods. *Karadeniz Fen Bilimleri Dergisi* 12 (2):916–30. doi:10.31466/KFBD.1174591.

Gutkin, M., R. Shamir, and G. Dror. Jul, 2009. SlimPLS: A method for feature selection in gene expression-based disease classification. *PLoS One* 4(7):e6416. doi:10.1371/JOURNAL.PONE.0006416.

Hou, J., Y. Sang, Y. Liu, and L. Lu, "Feature selection and prediction Model for type 2 diabetes in the Chinese Population with machine learning," *ACM International Conference Proceeding Series*, Oct. 2020, doi:10.1145/3424978.3425085.

Hsu, H. H., C. W. Hsieh, and M. Da Lu. Jul, 2011. Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications* 38 (7):8144–50. doi:10.1016/J.ESWA.2010.12.156.

Huerta, E. B., R. M. Caporal, M. A. Arjona, and J. C. H. Hernández. 2013. Recursive feature elimination based on linear discriminant analysis for molecular selection and classification of diseases. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7996:244–51. doi:10.1007/978-3-642-39482-9_28/COVER/.

Hu, M., and F. Wu. 2010. Filter-wrapper hybrid method on feature selection. *Proceedings - 2010 2nd WRI Global Congress on Intelligent Systems, GCIS 2010*, 3:98–101. doi:10.1109/GCIS.2010.235.

Jain, S., and A. Saha. Mar, 2022. Rank-based univariate feature selection methods on machine learning classifiers for code smell detection. *Evolutionary Intelligence* 15(1):609–38. doi:10.1007/s12065-020-00536-z.

Joshi, R. D., and C. K. Dhakal. Jul, 2021. Predicting type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health* 18(14). doi:10.3390/IJERPH18147346.

Juneja, A., S. Juneja, S. Kaur, and V. Kumar. 2021. Predicting Diabetes Mellitus With Machine Learning Techniques Using Multi-Criteria Decision Making. *International Journal of Information Retrieval Research* 11 (2):38–52. doi:10.4018/IJIRR.2021040103.

Kakoly, I. J., M. R. Hoque, and N. Hasan. 2023. Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique. *Sustainability* 15 (6):4930. doi:10.3390/SU15064930.

Kira, K., and L. A. Rendell. 1992. A practical approach to feature selection. *Machine Learning Proceedings* 1992 (Jan):249–56. doi:10.1016/B978-1-55860-247-2.50037-1.

Kishor, A., and C. Chakraborty. Jun, 2021. Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. *International Journal of Systems Assurance Engineering and Management* 1–9. doi:10.1007/s13198-021-01174-z.

Kulkarni, A., D. Chong, and F. A. Batarseh. Jan, 2020. Foundations of data imbalance and solutions for a data democracy. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering* 83–106. doi:10.1016/B978-0-12-818366-3.00005-8.

Kumari, S., D. Kumar, and M. Mittal. 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering* 2 (Jun):40–46. doi:10.1016/J.IJCCE.2021.01.001.

Li T, and Fong S. Nov, 2019. A fast feature selection method based on coefficient of variation for diabetics prediction using machine learning. *International Journal of Extreme Automation and Connectivity in Healthcare (IJEACH)* 1(1):55–65. doi:10.4018/IJEACH.2019010106.

Liu, Y., J. M. Wu, M. Avdeev, and S. Q. Shi. Feb, 2020. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties. *Advanced Theory and Simulations* 3(2):1900215. doi:10.1002/ADTS.201900215.

Madurapperumage, A., W. Y. C. Wang, and M. Michael. 2021. A systematic review on extracting predictors for forecasting complications of diabetes mellitus. In *ACM International Conference Proceeding Series*, May, 327–30. doi:10.1145/3472813.3473211.

Masoudi-Sobhanzadeh, Y., H. Motieghader, and A. Masoudi-Nejad. Apr, 2019. FeatureSelect: A software for feature selection based on machine learning approaches. *BMC Bioinformatics* 20(1):1–17. doi:10.1186/s12859-019-2754-0.

Mishra, S., H. K. Tripathy, P. K. Mallick, A. K. Bhoi, and P. Barsocchi. 2020. EAGA-MLP—an enhanced and adaptive hybrid classification Model for diabetes diagnosis. *Sensors* 20 (14):4036. doi:10.3390/S20144036.

Mucherino, A., P. J. Papajorgji, and P. M. Pardalos. 2009. Nearest neighbor classification. 83–106. doi:10.1007/978-0-387-88615-2_4.

Nagaraj, P., P. Deepalakshmi, R. F. Mansour, and A. Almazroa. 2021. Artificial flora algorithm-based feature selection with gradient boosted tree Model for diabetes classification. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* 14:2789–806. doi:10.2147/DMSO.S312787.

Oladimeji, O. O., A. Oladimeji, and O. Oladimeji. May, 2021. Classification models for likelihood prediction of diabetes at early stage using feature selection. *Applied Computing & Informatics* ahead-of-print. doi:10.1108/ACI-01-2021-0022.

Ottenbacher, K. J. Jul, 1995. The chi-square test: Its use in rehabilitation research. *Archives of Physical Medicine and Rehabilitation* 76 (7):678–81. doi:10.1016/S0003-9993(95)80639-3.

Papatheodorou, K., M. Banach, M. Edmonds, N. Papanas, and D. Papazoglou. 2015. Complications of diabetes. *Journal of Diabetes Research* 2015:1–5. doi:10.1155/2015/189525.

Pearson's Correlation Coefficient. 2008. *Encyclopedia of Public Health*. 1090–91. doi:10.1007/978-1-4020-5614-7_2569.

Pima Indians Diabetes Database | Kaggle. Accessed Jun 23, 2022. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Pirgazi, J., M. Alimoradi, T. Esmaeili Abharian, and M. H. Olyaee. Dec, 2019. An efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets. *Scientific Reports 2019* 9 (1):1–15. doi:10.1038/s41598-019-54987-1.

Ramesh, J., R. Aburukba, and A. Sagahyroon. Jun, 2021. A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters* 8 (3):45–57. doi:10.1049/HTL2.12010.

Ratner, B. Jun, 2009. The correlation coefficient: Its values range between 1/1, or do they. *Journal of Targeting Measurement & Analysis for Marketing* 17 (2):139–42. doi:10.1057/jt.2009.5.

Sabitha, E., and M. Durgadevi. 2022. Improving the diabetes Diagnosis prediction rate using data preprocessing, data augmentation and recursive feature elimination method. *IJACSA) International Journal of Advanced Computer Science and Applications* 13 (9). doi: 10.14569/IJACSA.2022.01309107.

Sahu, B., S. Dehuri, and A. Jagadev. Aug, 2018. A study on the relevance of feature selection methods in microarray data. *The Open Bioinformatics Journal* 11 (1):117–39. doi:10.2174/1875036201811010117.

Saxena, R., S. K. Sharma, M. Gupta, and G. C. Sampada. 2022. A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. *Computational Intelligence and Neuroscience* 2022:1–11. doi:10.1155/2022/3820360.

Sheik Abdullah, A., and S. Selvakumar. Oct, 2019. Assessment of the risk factors for type II diabetes using an improved combination of particle swarm optimization and decision trees by evaluation with Fisher's linear discriminant analysis. *Soft Computing* 23 (20):9995–10017. doi:10.1007/s00500-018-3555-5.

Sneha, N., and T. Gangil. 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*. doi:10.1186/s40537-019-0175-6.

Tadist, K., S. Najah, N. S. Nikolov, F. Mrabti, and A. Zahi. Dec, 2019. Feature selection methods and genomic big data: A systematic review. *Journal of Big Data* 6 (1):1–24. doi:10.1186/s40537-019-0241-0.

Tiwari, P., and V. Singh. Jan, 2021. Diabetes disease prediction using significant attribute selection and classification approach. *Journal of Physics Conference Series* 1714 (1):012013. doi:10.1088/1742-6596/1714/1/012013.

Tomic, D., J. E. Shaw, and D. J. Magliano. Sep, 2022. The burden and risks of emerging complications of diabetes mellitus. *Nature Reviews Endocrinology* 18 (9):525–39. doi:10.1038/S41574-022-00690-7.

Unnikrishnan, R., R. M. Anjana, and V. Mohan. 2016. Diabetes mellitus and its complications in India. *Nature Reviews Endocrinology* 12 (6):357–70. doi:10.1038/nrendo.2016.53.

Venkatesh, B., and J. Anuradha. 2019. A review of feature selection and its methods. *Cybernetics and Information Technologies* 19 (1):3–26. doi:10.2478/CAIT-2019-0001.

Yu, L., and H. Liu. Oct, 2004. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* 5:1205–24.

Zhang, T., T. Zhu, P. Xiong, H. Huo, Z. Tari, and W. Zhou. Mar, 2020. Correlated differential privacy: Feature selection in machine learning. *IEEE Transactions on Industrial Informatics / a Publication of the IEEE Industrial Electronics Society* 16 (3):2115–24. doi:10.1109/TII.2019.2936825.

Zhu, H., G. Liu, M. Zhou, Y. Xie, and Q. Kang. Jan, 2020. A noisy-sample-removed under-sampling scheme for imbalanced classification of public datasets. *IFAC-Papersonline* 53 (5):624–29. doi:10.1016/J.IFACOL.2021.04.202.