Improved Diabetes Prediction Using a Modified Linear Discriminant Algorithm

Amit Malik[1], Akshat Agrawal[2] and Zeba Khanam[3,*]

[1]Dept. of Computer Science & Engineering, SRM University Delhi-NCR, Haryana, India; amit.d@srmuniversitya.c.in
[2] Dept. of Computer Science & Engineering, Amity University Haryana, Gurugram India; akshatag20@gmail.com
[3] College of Computing and Informatics, Saudi Electronic University, z.khanam@seu.edu.sa

***Corresponding Author:** Zeba Khanam, z.khanam@seu.edu.sa

Abstract

Retinopathy is damage to the retina of the eye, a serious microvascular complication of diabetes. CAD tools for DR management are typically developed using AI methods like machine learning and deep learning algorithms. In recent times, diagnostic instruments for DR have been developed employing deep learning models. Due to this fact, a large amount of data is needed for these models' training. The enormous amounts of data are uneven since there are fewer instances in the dataset. This research introduces a new paradigm called modified LDA, which uses small amounts of training data to successfully train models, thereby avoiding the common problems of overfitting and approximation error that come when working with such limited data. Through our study, we present a novel method—a modified linear discriminant algorithm—for categorizing and diagnosing diabetic patients. Information gathered from the Kaggle. Accuracy (97.92%), area under the curve (0.9999), and Gini index (0.998) are all rising numbers. By analyzing objective performance measures and the interpreted model, we find that the suggested model outperforms the state-of-the-art methods in identifying individuals with diabetes and ranking the severity of their disease when missing values are present. Thus, an ophthalmologist may be able to receive a second opinion on the severity of the diabetic illness with the help of this tool.

Keywords

Diabetes prediction, linear discriminant algorithm, modified algorithm, machine learning, predictive modeling, classification, feature selection, data analysis

1. Introduction

Scientists believe that an immunological response leads to type 1 diabetes (the body attacks itself by mistake). Caused by this response, insulin production is halted. Type 1 diabetes is characterised by rapid onset of symptoms. Children, teenagers, and young adults are the typical patients. Healthy lifestyle modifications, such as those listed below, can help prevent or postpone the onset of type 2 diabetes.

- Dropping pounds.
- Healthy eating habits.
- Involved in physical activity.

Type 2 diabetes symptoms frequently manifest gradually over time. It is possible for a person to have type 2 diabetes for many years without ever being diagnosed. In the event that symptoms do arise, they may include:

- Dryness of the mouth
- Lack of control over urination frequency
- Hunger has increased.
- Dropping pounds without trying
- Fatigue
- Inability to focus clearly
- Bruises that take a long time to heal
- Repeated infections
- Tingling or numbness in the extremities
- Discolored patches of skin

Type 2 diabetes is mostly caused by one's lifestyle and typically manifests later in life, while type 1 diabetes is primarily tied to one's genes and typically manifests in childhood.

Gestational Diabetes

Diabetes mellitus occurs during pregnancy in women who have never had the disease before. Baby's health is at stake if mom has gestational diabetes. After giving birth, most women with gestational diabetes no longer have symptoms [3]. But it raises the probability that you'll get type 2 diabetes down the road. In the long run, your child is at a higher risk of being overweight or obese. People who already have health problems tend to suffer more from coronavirus infection. Therefore, it is well-documented that diabetics have a greater susceptibility to infection with corona virus.

Diabetes can lead to long-term complications affecting several body systems, including the cardiovascular system, blood vessels, eyes, kidneys, and nerves[4].

- An key factor in preventable blindness, diabetic retinopathy is the result of progressive damage to the retina's tiny blood vessels brought on by long-term diabetes. Because of diabetes, about a million people are visually impaired (2).
- One of the most common reasons for renal failure is diabetes (3).

Type 1 diabetes is a hereditary illness that often manifests at an early age, while type 2 diabetes is connected mostly to lifestyle factors and typically appears later in life. Type 1 diabetes cannot be avoided in its early stages in any known method. On the other hand, scientists are researching on ways to protect newly diagnosed individuals' islet cells from becoming irreparably damaged.

Prevention

Although there is currently no treatment for type 2 diabetes, our researchers are making progress on a groundbreaking trial using the use of weight control to assist patients achieve remission. When glucose (or sugar) levels in the blood return to normal, the disease is said to be in remission. Unfortunately, this does not spell the end for diabetes. It is nevertheless crucial for persons in remission to maintain frequent checkups with their doctors. On the other hand, achieving remission can drastically alter one's outlook on life.Preventing or postponing the onset of type 2 diabetes can be achieved by the adoption of a healthy lifestyle [5]. Type 2 diabetes and its consequences can be avoided if persons take the following measures:

- reach and keep a normal weight;
- engage in moderateintensity physical exercise for at least 30 minutes on most days. In order to maintain a healthy w eight, you need to be more active.
- consume a balanced diet low in sugar and saturated fats; and
- Stay away from tobacco products; doing so lowers your protection against diabetes and cardiovascular disease.

Diagnostics and therapy

The fasting blood glucose test evaluates glucose levels in the blood after a fast of at least eight hours (not eating). Testing blood sugar is cheap and can be done to make a diagnosis early. Reducing blood glucose and other risk factors that damage blood arteries is an important part of treating diabetes. Fasting blood glucose ≥ 126 mg/dl is diagnostic threshold for diabetes. Quitting smoking is also recommended to reduce the risk of negative outcomes [6-10].

Their blood sugar is constantly monitored by a sensor implanted beneath their skin. Wearables and mobile phones can receive data via a transmitter. Among the most effective and least expensive interventions for nations with limited resources are:

- helping those with type 2 diabetes, and insulin may be necessary in some cases;
- managing one's blood pressure;

Problems with the heart, kidneys, feet, mouth, eyes, and mind are all common consequences of diabetes. Various other cost-cutting measures include:

- Screening for and treatment of blindness-causing retinopathy;
- Management of blood lipids (for cholesterol control);
- Early detection and management of diabetic kidney disease by screening.

Humanity has made tremendous strides in the fields of computer science, material science, biotechnology, genomics, and proteomics in recent years. BP < 130/80, Trig <150, LDL < 100 are the recommended blood pressure and lipid goals for the prevention of cardiovascular disease in adults with diabetes. The status quo in healthcare is changing due to these innovative technology. Particularly, AI and big data are altering illness and patient care, with the emphasis moving toward individualised diagnosis and therapy. Because of this change, public health may now focus on prevention and prediction. Hence these facts motivate the authors to propose this model to predict multiple stages of the diabetic disease.

The rest of this work is organized as follows: Sect. 2 describes the details of the existing works; in Sect. 3, the proposed methods are utilized to predict the Diabetes disease; the results are being represented in section 4 and finally, the conclusions are summarized.

2. Related Work

The use of machine learning in diabetes has been the subject of previous evaluations, but these analyses took a very different tack.

Gathering the mistake remedy learning, the back likelihood dispersion of loads given the blunder capability, and the Goodman-Kruskal Gamma rank relationship into a Bayesian learning approach is proposed in another preparation technique by Belciug et al. [11]. The essential objectives of this examination were (1) to make another learning technique that integrates the Bayesian worldview and the mistake back-engendering, and (2) to assess the progress of this strategy. Measurements show that the new technique for advancing reliably beats the old ones.

With the use of social media data mining, Lim et al. [12] offer an unsupervised machine learning model that may detect hidden viral illnesses in the real world. Most current methods for simulating knowledge discovery about infectious illnesses via social media networks take a top-down approach, building on what is already known about the topic at hand (e.g., disease names and symptoms). Latent infectious illness formalisation procedures tend to be laborious and drawn out. Therefore, this exploration presents a base up methodology for distinguishing idle irresistible sicknesses in a particular region without earlier information, for example, disease names and related side effects.

In the wake of thinking about the presence of mental comorbidities, Marrie et al. [13] look at the connection among diabetes and hypertension in individuals with MS and their mental capability. A singular's z-score subsequent to adapting to mature, orientation, and level of training might be seen with regards to the entire test-taking populace. Utilizing a multivariate direct model, we inspected the connection among diabetes and hypertension and the four mental z-scores, controlling for the presence of other psychological well-being conditions, for example, melancholy and uneasiness as well as the utilization of any psychotropic medications, illness changing treatments, smoking history, and weight file. Most of the 111 people were female (82.9%) and had backsliding transmitting MS (83.5%). The typical age of the members was 49.6 years (SD: 12.7). 22.7% of patients had hypertension, 10.8% had diabetes, 9.9% had serious sadness, and 9.9% had tension issues.

Elhadd et al. [14] make a machine-based calculation using clinical and segment information, active work, and glucose fluctuation. 13 patients (10 men and 3 females) with type 2 diabetes who were using three or more anti-diabetic drugs were monitored for two weeks before to Ramadan and for two weeks throughout the holy month. Using a regression framework informed by past and present levels of physical activity and blood sugar, many machine learning methods were educated to make predictions about future blood sugar levels during Ramadan and other times of the year. The average age of the individuals was 51 (interquartile range [IQR] 49-52), their average body mass index was 33.2 (IQR 33-35.9), and their average haemoglobin A1c was 7.3%. (IQR 6.7–7.8). R2 for the most effective model incorporating physical activity was 0.548, with an MAE of 30.30.

Papers with the word "prediction" in the title were analysed by Varga et al. [15] in the field of diabetes research. We checked to see if these publications had metrics helpful for making forecasts. NCBI PubMed was used for a comprehensive search. Diabetes-related and forecasting-related articles were chosen. Predictive metrics were sought out by searching the abstracts of all original research publications published in the field of diabetes epidemiology. There were 2,182 results found matching your query. It was determined that 1,910 abstracts were worth reviewing after removing irrelevant papers. Just under four-fifths (n = 745) of these reported using predictive statistical measures, whereas almost two-thirds (n = 1,165) did not. Prediction measures such as ROC AUC, sensitivity, and specificity were most commonly provided.

Risk factors for death among hospitalised adults with diabetes are identified by Mukherjee et al. [16]. The PRISMA protocol served as the foundation for the systematic review. EMBASE and MEDLINE were searched for relevant studies. Research articles published in English during the previous six years that examined post-hospital mortality risk factors in adult patients with diabetes were considered for inclusion. The next step was a "quality evaluation and semi-quantitative synthesis in accordance with PRISMA principles" of the extracted data. There were 35 studies that met the criteria for inclusion, all of which looked at mortality risk factors for diabetic hospital dischargees. These reports go all over the world. There are 48 different risk factors for death that have been found to have a substantial statistical impact. The test findings, and glycaemic state are all examples of possible classes of risk factors.

According to Nnamoko et al. [17], one of the biggest challenges for machine learning classifiers is learning from outliers and unbalanced data. Data preparation solutions are well-known for being effective and simple to deploy among the several methods designed to address this issue. To ensure an equal distribution, we offer a selective data preparation strategy that incorporates prior knowledge about outlier occurrences into a synthetically created subset. To achieve diversity in the training data, we employed the Synthetic Minority Oversampling TEchnique (SMOTE). First, though, the outliers had to be singled out and oversampled (irre- spective of class). The goal is to prevent outliers from having too much of an impact on the training dataset. The experiments demonstrate that SMOTE is enhanced by such judicious oversampling, which in turn leads to better classification results.

To find stowed away disease groups and patient subgroups from EHRs, Wang et al. [18] take a gander at the utilization of unaided AI models. To find dormant illness groups and patient subgroups, we led exact tests looking at LDA and PDM on EHR information from the Rochester Epidemiology Project (REP) clinical records linkage framework. Utilizing a graphical portrayal of sicknesses, we analyzed the productivity of LDA and PDM in finding disease groups.

A climb in plantar temperature not long before DFU might be discernible with the utilization of thermogram pictures, as portrayed by Khandakar et al. [19]. In any case, in light of the fact that plantar temperature may not be consistently circulated, it very well may be trying to measure and use in prescient models. To distinguish the diabetic foot, we have tried many cutting edge Convolutional Neural Networks (CNNs) utilizing foot thermogram pictures and have proposed a suitable arrangement in view of an AI based scoring approach with highlight choice and enhancement methods and learning classifiers. The AdaBoost Classifier with 10 elements created a F1 score of 97%, while the shallower CNN model MobilenetV2 scored 95% on a two-foot thermogram picture based grouping. The proposed procedure might be executed as a cell phone application to permit the client to watch the improvement of the DFU at home, as confirmed by a correlation of the derivation time for the best-performing organizations.

By distinguishing diabetics and prediabetics at a beginning phase and mediating at the fitting time, Li et al. [20] can stay away from the headway of prediabetics to diabetics and delay the movement to diabetes, which has positive ramifications for general wellbeing. We make the painless diabetics risk expectation model in light of tongue qualities combination utilizing AI procedures, and we get pictures of the tongue, extricate variety and surface elements utilizing TDAS, and remove progressed highlights utilizing ResNet-50. We then accomplish the combination of these two arrangements of highlights utilizing GA XGBT, lay out a harmless diabetics risk expectation model, and evaluate the testing's viability. The GA XGBT model with combination highlights has the most elevated, as indicated by the cross-approval. With a typical CA of 0.81, an AUROC of 0.918, an AUPRC of 0.839, a Precision of 0.821, a Recall of 0.81, and a F1-score of 0.796 on the test set, the GA XGBT model seems to perform best.

Utilizing normal language handling (NLP) and AI, Brown et al. [21] planned to foster novel, legitimate, and adaptable appraisals of patients' wellbeing education (HL) and doctors' semantic intricacy (ML). We utilized these techniques to over

400k patient-specialist SMs sent and got through an electronic patient entrance, and we made and approved a robotized patient education profile (LP) and doctor intricacy profile (CP). Here, we talk about the obstructions that must be survived and the strategies that were utilized to conquer them all through this notable exertion. Both the authority concentrate on records and meetings with the analysts were used to portray the issues and their cures. The group utilized a mix of Google Docs capacities and a web-based group coordinating, following, and the board instrument to monitor their examination over the span of the five-year project (Asana). The group met a few times all through year 5 to distinguish, order, and code the most squeezing issues and their separate arrangements. From three expansive cycle spaces, we distinguished 23 issues and related strategies.
DFUC2020's discoveries are summed up by Hoon et al. [22], who assess the profound learning-based calculations presented by the triumphant groups. We portray the model engineering, preparing boundaries, and additional stages (pre-handling, information expansion, and post-handling) for every profound learning approach exhaustively. Our investigation is intensive, including the qualities and shortcomings of every strategy. To prepare on more photographs, each of the methodologies required an information expansion of some kind or another, and misleading up-sides must be sifted through here and there during post-handling. At last, we show that a group approach using a few profound learning calculations might further develop it.
Clinical results of rivaroxaban and warfarin in people with nonvalvular atrial fibrillation (NVAF) and heftiness/diabetes are looked at by Weir et al. [23]. Patients beyond 18 years old who met the accompanying rules were chosen from a medical care claims information base: new inception of rivaroxaban or warfarin, something like one clinical case with a finding of AF, still up in the air by an approved AI calculation, and no less than one clinical case with a conclusion of diabetes or for antidiabetic medicine. Cox corresponding perils models were utilized to analyze stroke/foundational embolism (SE) and significant draining rates between treatment bunches that were matched utilizing inclination scores. Patients with non-valvular atrial fibrillation (NVAF) who were overweight and had diabetes and begun on rivaroxaban or warfarin were coordinated. The gamble of serious draining didn't vary fundamentally across treatment gatherings (HR 0.92, 95% CI 0.78-1.09).
The majority of the work of Murugappan et al. [24] was created with the use of AI techniques including machine and deep learning algorithms. Deep learning models have been used to provide diagnostic tools for DR in recent years. Consequently, a lot of data is needed to properly train these models. Because there are fewer examples in the dataset, the massive volumes of data are unbalanced. In this research, it leverages a very limited amount of training data to train the models successfully, therefore avoiding the issues of overfitting and poor approximation that arise when training models with tiny datasets. In order to grade and identify DR based on attention, this research introduces a unique prototype network, a kind of FSL classification network. Few-shot classification problems are ideal for the DRNet framework's episodic learning model training methodology. For the purpose of diagnosing and classifying diabetic patients, we created a DRNet using the APTOS2019 dataset. As a means of capturing visual representations, the suggested network aggregates transformations and uses class gradient activations to construct the attention mechanism.
Long-term problems, such as those caused by Gollapalli et al. [25], reduce one's quality of life. These include, among others, blindness, renal failure, and heart disease. There has been a tenfold increase in reported cases of diabetes in Saudi Arabia during the past three years. Three main types of diabetes mellitus (DM) have been identified. Because of this uncertainty over the accurate diagnosis, doctors often have a hard time keeping patients' conditions under control as they worsen. Attempts to foresee the onset of type 2 diabetes have been attempted in earnest. However, research on the most effective methods of diagnosing type 1 diabetes and prediabetes is lacking. They were utilized all through four tests determined to accomplish the most ideal results. SMOTE was utilized to accomplish information equality in Experiments 2, 3, and 4. Observational outcomes showed that the imaginative Utilizing change highlight importance, we tracked down that five essential elements, including Education, significantly influenced the model's exactness.

3. Materials and Methods

Dataset

Several machine learning methods have been run on the Diabetes dataset for this work. The authors make forecasts with this data collection. Many factors are measured and recorded in this data set, including Body Mass Index, Family History of Diabetes, Age, and Outcome [26]. The following is the download URL for the dataset:
https://www.kaggle.com/datasets/pritsheta/diabetes-dataset

Data preprocessing

Data pre-processing, also known as data wrangling, is the process of preparing data for analysis by applying various techniques (such as importing libraries, data, checking of missing values, categorical following validation and feature scaling) to transform raw data (which may be incomplete, inconsistent, error-ridden, or lacking in certain behaviour) into a usable format [27]. The model is trained using a wide variety of acquired data. Raw data is data that has not been processed in any way; it may have errors or outliers; important information may be presented in a string format; numerical values may be presented in a string format; etc. Both the speed and precision of machine learning models may be improved with careful data pre-processing. Because it aids in cleaning up a dataset and revealing its true significance.

Handling missing values

It is possible that there will be blanks in the feature matrix once it is constructed. It might be an issue during training if the authors don't take care of it now. The imputation strategy finds plausible replacements for missing information. When a little

fraction of the data is absent, it excels. If there is a significant amount of blanks, it will be impossible to get reliable findings. The alternative is to delete information. The elimination of correlated data can help mitigate the effects of missing data that occurs at random. If there aren't enough observations for a trustworthy analysis, removing data may not be the best solution. There are times when keeping an eye out for a certain set of circumstances is essential.
Two strategies are available for dealing with the missing data:

- Getting rid of the whole row that includes the erroneous value works, but you can lose some important data in the process. If the dataset is substantial, this method may be useful.
- If a number in a column is missing, you may get a rough idea of what it should be by calculating its mean, median, mode, etc.

Handling outliers

Data points that deviate wildly from the norm are said to be outliers. As a result of uneven data input or erroneous observations, outliers can occur and distort the data. Outliers must be identified and removed to guarantee that the trained model can accurately generalise to the valid range of test inputs [28]. The authors have used the data's standard deviation, or an equivalent z-score, to identify outliers when the data, or a subset of characteristics within the dataset, follows a normal distribution. The standard deviation is a statistical metric used to quantify how distant individual data points are from the mean.

Linear Discriminant Analysis

To classify data and reduce the number of dimensions, Linear Discriminant Analysis uses a linear model. Most frequently encountered in pattern classification tasks when feature extraction is required. As you can see, this has been in place for quite some time. LDA minimises variability within classes while maximising variability between them [29].
The failures of Logistic Regression are necessary background reading for grasping why LDA was developed. The logistic regression method has the following restrictions:

- Classically, the domain of Logistic Regression has been binary classification issues. In theory, it might be generalised and applied to multi-class classification tasks, but in practise, this is rarely done.
- When the classes are clearly differentiated, the stability of Logistic Regression may suffer. When there is a large gap between socioeconomic groups, therefore, there is instability.
- When there are few characteristics to estimate parameters for, Logistic Regression becomes unstable.

The aforementioned flaws in logistic regression have been addressed with modified LDA. The predicator importance has been displayed in the figure 1 as given below:

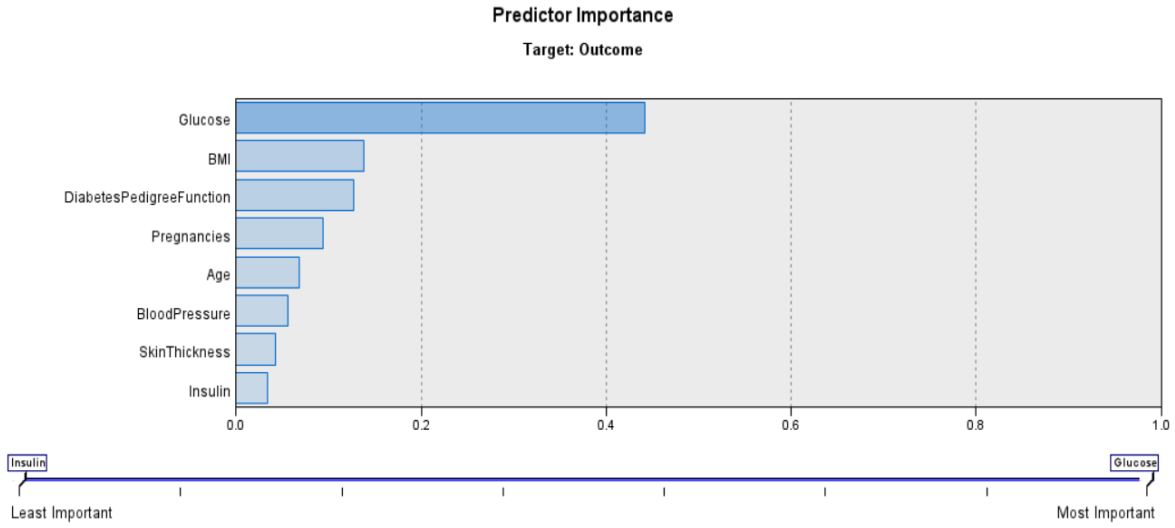

Figure 1 Predictor Importance

Classifying data into binary and non-binary categories, modified Linear Discriminant Analysis use a linear method to discover the connection between dependent and independent characteristics. By applying the Fischer formula, the data dimensions are flattened such that they may be accommodated in a linear space. Modified LDA is an efficient classifier, dimensionality reduction, and data visualizer all in one [30]. The goal of modified LDA is to:

- To reduce variability between classes, by grouping together as many comparable data points as feasible. More accurate categorizations are achieved in this way.
- For maximum confidence in the predictions made, the mean is put as far as feasible from the centre of each class.

Steps in predicting the diabetics through modified Linear Discriminant Analysis

Step 1: Computing the d-dimensional mean vectors
For a vector space to be considered "finite-dimensional," its basis must be made up of a finitely small set of vectors. Due to the fact that all bases of a finite-dimensional vector space have the same amount of elements, this is the dimension of the space.

Step 2: Computing the Scatter Matrices

- Within-class scatter matrix S_W

Table 1 Group Statistics

Outcome		Valid N (listwise)	
		Unweighted	Weighted
1	Pregnancies	234	234.000
	Glucose	234	234.000
	BloodPressure	234	234.000
	SkinThickness	234	234.000
	Insulin	234	234.000
	BMI	234	234.000
	DiabetesPedigreeFunction	234	234.000
	Age	234	234.000
0	Pregnancies	452	452.000
	Glucose	452	452.000
	BloodPressure	452	452.000
	SkinThickness	452	452.000
	Insulin	452	452.000
	BMI	452	452.000
	DiabetesPedigreeFunction	452	452.000
	Age	452	452.000
Total	Pregnancies	686	686.000
	Glucose	686	686.000
	BloodPressure	686	686.000
	SkinThickness	686	686.000
	Insulin	686	686.000
	BMI	686	686.000
	DiabetesPedigreeFunction	686	686.000
	Age	686	686.000

- Between-class scatter matrix S_B

Table 2 Functions at Group Centroids

Outcome	Function
	1
1	.950
0	-.492

Step 3: Solving the generalized eigenvalue problem for the matrix $S^{-1}_W SB$

- Checking the eigenvector-eigenvalue calculation

Table 3 Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.469[a]	100.0	100.0	.565

a. First 1 canonical discriminant functions were used in the analysis.

Step 4: Selecting linear discriminants for the new feature subspace

- Sorting the eigenvectors by decreasing eigenvalues

Table 4 Standardized Canonical Discriminant Function Coefficients

	Function
	1
Pregnancies	.321
Glucose	.772

BloodPressure	-.207
SkinThickness	.023
Insulin	-.079
BMI	.408
DiabetesPedigreeFunction	.211
Age	.148

- Choosing k eigenvectors with the largest eigenvalues

Table 5 Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.681	261.468	8	.000

Step 5: Transforming the samples onto the new subspace

Table 6Structure Matrix

	Function
	1
Glucose	.812
BMI	.442
Age	.378
Pregnancies	.356
DiabetesPedigreeFunction	.257
Insulin	.202
SkinThickness	.119
BloodPressure	.093

4. Results and Discussion

Experimental setup

In this work, this modified algorithm has been implemented in Python language on the diabetics dataset [31]. Below the table 7 is showing the settings selected for the experiment by authors in Python during executing the proposed model. The Defined Parameters represent the high-level parameters. These Accuracy, Time, and Interpretability settings map to the following internal configuration of this experiment

Table 7 System Parameters

Defined Parameter	**Value**
accuracy	7
time	2
interpretability	8
num_prediction_periods	None
num_gap_periods	None
time column	[OFF]

The data is being split into 70% for training and 30% for testing purpose. The figure 2 depicts the results obtained by executing the proposed model on prescribed dataset as given below:

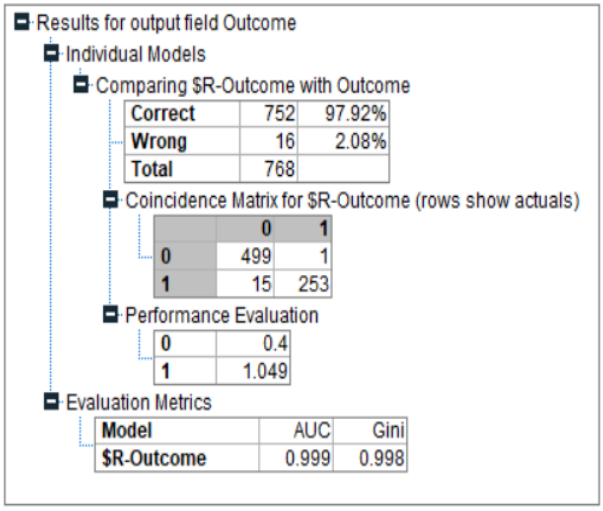

Figure 2 Results obtained in Modified LDA

The accuracy level in this algorithm is 97.92% in prediction of diabetics disease [32]. The AUC & Gini value in this algorithm is 0.99 and 0.998 respectively. To determine the Gini Index, sometimes called the Gini Impurity, we take the square root of the difference between the total squared probability for each class and one. It's easiest to apply and works best with larger sections.

The Gini Index can take on values between 0 and 1, with 0 indicating perfect classification accuracy and 1 indicating completely random assignment of data to categories. If the Gini Index is 0.5, then the items in certain classes are evenly distributed. Table 8 depicts the he accuracy achieved by various machine learning algorithms and the proposed model as given below [33].

Table 8 Performance Comparison

S. No.	Algorithm	Accuracy	AUC
1	SVM	76.823	0.839
2	CHAID	77.604	0.860
3	C&RT	77.734	0.741
4	Logistic Regression	78.255	0.839
5	Neural Network	78.646	0.859
6	Bayesian Network	81.25	0.878
7	C5.0	83.984	0.889
8	Random Tree	92.578	0.982
9	XGBoost	92.969	0.980
10	Proposed Model	97.92	0.998

The results demonstrated that there is no variance faced in the execution of the proposed algorithm in predicting the diabetics among different age of the patients [34]. Figure 3 demonstrated the accuracy comparison between various machine learning algorithms and modified LDA as given below:

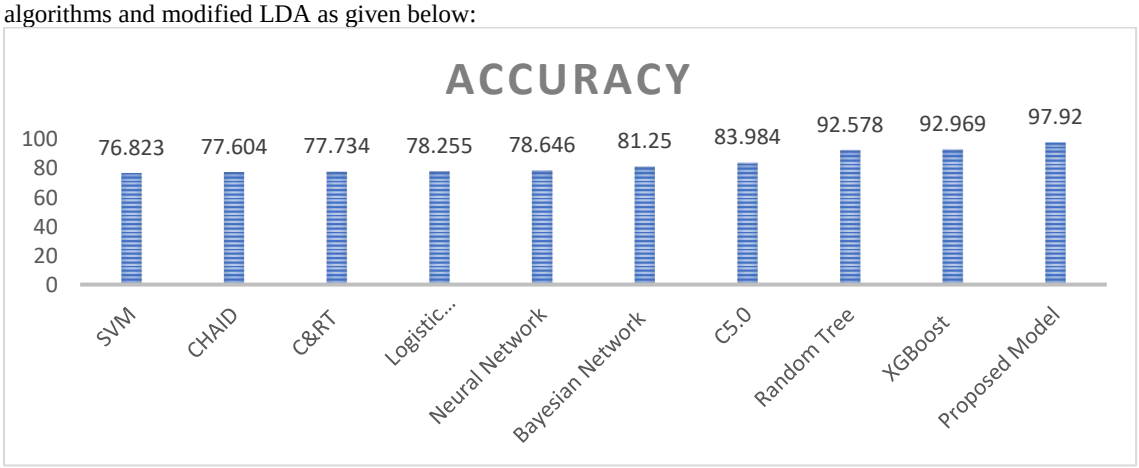

Figure 3 Results comparison

Regular urination, increased thirst, and persistent hunger are all signs of elevated blood sugar. When diabetes isn't addressed, it can lead to a host of issues. Diabetic ketoacidosis, hyperosmolar hyperglycemic condition, and mortality are all possible acute consequences. The prosed model helps the patient to prevent all serious long-term consequences i.e. Heart disease, stroke, chronic renal disease, foot ulcers, and eye impairment [35].

5. Conclusions

Cells, the fundamental units of life, primarily utilise glucose for energy. Insulin provides it to the body, where it is used for vital metabolic processes. Diabetes mellitus (DM), a prevalent kind of chronic illness, is characterised by an abnormality in glucose levels. Long-term effects include impaired vision, renal failure, and cardiovascular disease, all of which reduce an individual's standard of living. Diabetic micronutrient deficiencies are a common complication of all three types of diabetes mellitus. It can be challenging for doctors to manage the course of a disease when they are uncertain about the precise diagnosis. Predicting type 2 diabetes has been a major research focus. However, research on the most effective methods of diagnosing type 1 diabetes and prediabetes is lacking. Therefore, the purpose of this research is to employ Machine Learning (ML) to differentiate and forecast the three forms of diabetes using data from the above mentioned. We used the LDA to achieve data parity in Experiments.

References

[1]. Kaur, Manjit, and Dilbag Singh. "Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks." Journal of Ambient Intelligence and Humanized Computing 12, no. 2 (2021): 2483-2493.

[2]. S. Kim, H. Kim, E. Lee, C. Lim, and J. Lee, “Risk score-embedded deep learning for biological age estimation: Development and validation,” Inf. Sci. (Ny)., vol. 586, pp. 628–643, 2022, doi: 10.1016/j.ins.2021.12.015.

[3]. T. Mukherjee et al., “Journal of Diabetes and Its Complications A systematic review considering risk factors for mortality of patients discharged from hospital with a diagnosis of diabetes,” J. Diabetes Complications, vol. 34, no. 11, p. 107705, 2020, doi: 10.1016/j.jdiacomp.2020.107705.

[4]. Hooda, M., & Shravankumar Bachu, P. (2020). Artificial Intelligence Technique for Detecting Bone Irregularity Using Fastai. In International Conference on Industrial Engineering and Operations Management Dubai, UAE (pp. 2392-2399).

[5]. Arora, S., & Dalal, S. (2019). An optimized cloud architecture for integrity verification. Journal of Computational and Theoretical Nanoscience, 16(12), 5067-5072.

[6]. Arora, S., & Dalal, S. (2019). Trust Evaluation Factors in Cloud Computing with Open Stack. Journal of Computational and Theoretical Nanoscience, 16(12), 5073-5077.

[7]. Shakti Arora, S. (2019). DDoS Attacks Simulation in Cloud Computing Environment. International Journal of Innovative Technology and Exploring Engineering, 9(1), 414-417.

[8]. Shakti Arora, S. (2019). Integrity Verification Mechanisms Adopted in Cloud Environment. International Journal of Engineering and Advanced Technology (IJEAT), 8, 1713-1717.

[9]. Sudha, B., Dalal, S., & Srinivasan, K. (2019). Early Detection of Glaucoma Disease in Retinal Fundus Images Using Spatial FCM with Level Set Segmentation. International Journal of Engineering and Advanced Technology (IJEAT), 8(5C), 1342-1349.

[10]. Sikri, A., Dalal, S., Singh, N. P., & Le, D. N. (2019). Mapping of e-Wallets With Features. Cyber Security in Parallel and Distributed Computing: Concepts, Techniques, Applications and Case Studies, 245-261.

[11]. Seth, B., Dalal, S., & Kumar, R. (2019). Hybrid homomorphic encryption scheme for secure cloud data storage. In Recent Advances in Computational Intelligence (pp. 71-92). Springer, Cham.

[12]. Seth, B., Dalal, S., & Kumar, R. (2019). Securing bioinformatics cloud for big data: Budding buzzword or a glance of the future. In Recent advances in computational intelligence (pp. 121-147). Springer, Cham.

[13]. Jindal, U., & Dalal, S. (2019). A hybrid approach to authentication of signature using DTSVM. In Emerging Trends in Expert Applications and Security (pp. 327-335). Springer, Singapore.

[14]. Le, D. N., Seth, B., & Dalal, S. (2018). A hybrid approach of secret sharing with fragmentation and encryption in cloud environment for securing outsourced medical database: a revolutionary approach. Journal of Cyber Security and Mobility, 7(4), 379-408.

[15]. Sikri, A., Dalal, S., Singh, N. P., & Dahiya, N. (2018). Data Mining and its Various Concepts. Kalpa Publications in Engineering, 2, 95-102.

[16]. W. Brown et al., “Challenges and solutions to employing natural language processing and machine learning to measure patients ’ health literacy and physician writing complexity : The ECLIPPSE study,” J. Biomed. Inform., vol. 113, no. December 2020, p. 103658, 2021, doi: 10.1016/j.jbi.2020.103658.

[17]. Dalal, S., Poongodi, M., Lilhore, U. K., Dahan, F., Vaiyapuri, T., Keshta, I., ... & Simaiya, S. Optimized LightGBM model for security and privacy issues in cyber-physical systems. Transactions on Emerging Telecommunications Technologies, e4771.

[18]. Dalal, S., Manoharan, P., Lilhore, U. K., Seth, B., Simaiya, S., Hamdi, M., & Raahemifar, K. (2023). Extremely boosted neural network for more accurate multi-stage Cyber attack prediction in cloud computing environment. Journal of Cloud Computing, 12(1), 1-22.

[19]. Shetty, S., & Dalal, S. (2022, December). Bi-Directional Long Short-Term Memory Neural Networks for Music Composition. In 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (pp. 1-6). IEEE.

[20]. Dalal, S., Seth, B., Radulescu, M., Cilan, T. F., & Serbanescu, L. (2023). Optimized Deep Learning with Learning without Forgetting (LwF) for Weather Classification for Sustainable Transportation and Traffic Safety. Sustainability, 15(7), 6070.

[21]. Onyema, E. M., Lilhore, U. K., Saurabh, P., Dalal, S., Nwaeze, A. S., Chijindu, A. T., ... & Simaiya, S. (2023). Evaluation of IoT-Enabled hybrid model for genome sequence analysis of patients in healthcare 4.0. Measurement: Sensors, 26, 100679.

[22]. Dalal, S., Manoharan, P., Lilhore, U. K., Seth, B., Simaiya, S., Hamdi, M., & Raahemifar, K. (2023). Extremely boosted neural network for more accurate multi-stage Cyber attack prediction in cloud computing environment. Journal of Cloud Computing, 12(1), 1-22.

[23]. Dalal, S., Goel, P., Onyema, E. M., Alharbi, A., Mahmoud, A., Algarni, M. A., & Awal, H. (2023). Application of Machine Learning for Cardiovascular Disease Risk Prediction. Computational Intelligence and Neuroscience, 2023.

[24]. Dalal, S., Seth, B., Radulescu, M., Secara, C., & Tolea, C. (2022). Predicting Fraud in Financial Payment Services through Optimized Hyper-Parameter-Tuned XGBoost Model. Mathematics, 10(24), 4679.

[25]. Dalal, S., Onyema, E. M., & Malik, A. (2022). Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy. World Journal of Gastroenterology, 28(46), 6551-6563.

[26]. Edeh, M. O., Dalal, S., Obagbuwa, I. C., Prasad, B. V. V., Ninoria, S. Z., Wajid, M. A., & Adesina, A. O. (2022). Bootstrapping random forest and CHAID for prediction of white spot disease among shrimp farmers. Scientific Reports, 12(1), 1-12.

[27]. Zaki, J., Nayyar, A., Dalal, S., & Ali, Z. H. (2022). House price prediction using hedonic pricing model and machine learning techniques. Concurrency and Computation: Practice and Experience, 34(27), e7342.

[28]. Dalal, S., Onyema, E., Romero, C., Ndufeiya-Kumasi, L., Maryann, D., Nnedimkpa, A. & Bhatia, T. (2022). Machine learning-based forecasting of potability of drinking water through adaptive boosting model. Open Chemistry, 20(1), 816-828. https://doi.org/10.1515/chem-2022-0187

[29]. Onyema, E. M., Dalal, S., Romero, C. A. T., Seth, B., Young, P., & Wajid, M. A. (2022). Design of Intrusion Detection System based on Cyborg intelligence for security of Cloud Network Traffic of Smart Cities. Journal of Cloud Computing, 11(1), 1-20.

[30]. Dalal, S., Onyema, E. M., Kumar, P., Maryann, D. C., Roselyn, A. O., & Obichili, M. I. (2022). A Hybrid machine learning model for timely prediction of breast cancer. International Journal of Modeling, Simulation, and Scientific Computing, 2023, 1-21.

[31]. Dalal, S., Seth, B., Jaglan, V., Malik, M., Dahiya, N., Rani, U., ... & Hu, Y. C. (2022). An adaptive traffic routing approach toward load balancing and congestion control in Cloud–MANET ad hoc networks. Soft Computing, 26(11), 5377-5388.

[32]. S. Belciug and F. Gorunescu, "Error-correction learning for artificial neural networks using the Bayesian paradigm . Application to automated medical diagnosis," J. Biomed. Inform., vol. 52, pp. 329–337, 2014, doi: 10.1016/j.jbi.2014.07.013.

[33]. R. Venugopal et al., "Privacy preserving Generative Adversarial Networks to model Electronic Health Records," Neural Networks, vol. 153, pp. 339–348, 2022, doi: 10.1016/j.neunet.2022.06.022.

[34]. D. Singh, M. Kaur, M. Y. Jabarulla, V. Kumar and H. -N. Lee, "Evolving fusion-based visibility restoration model for hazy remote sensing images using dynamic differential evolution," in IEEE Transactions on Geoscience and Remote Sensing., doi: 10.1109/TGRS.2022.3155765.

[35]. I. Czarnowski, "Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams," J. Comput. Sci., vol. 61, no. January, p. 101614, 2022, doi: 10.1016/j.jocs.2022.101614.

[36]. Sethi, N. Jaglan, V. Bhaskar, S. (2021) Multi-Operator based Saliency Detection. 2021 International Conference on Computational Performance Evaluation (ComPE), 507-512.

[37]. Swati, Jaglan, V. Bhaskar, S. (2021) A Novel Multi Granularity Locking Scheme Based On Concurrent Multi-Version Hierarchical Structure. Information Technology In Industry, 9(1), 932-947.

[38]. Sethi, N. Verma, J.K. Shrivastava, U. Bhaskar, S. (2021) Google Stock Movement: A Time Series Study Using LSTM Network. Multidisciplinary Functions of Blockchain Technology in AI and IoT Applications, 70-87.

[39]. Chaudhary, S. Jatain, A. Nagpal, P. Bhaskar, S. (2021) Design and Development of Gesture Based Gaming Console. A & V Publications, 12(2), 51-56.

[40]. Nagpal, P. Jatain, A. Chaudhary, S. Bhaskar, S. (2021) Motion detection of webcam using frame differencing method. A & V Publications, 12(2), 32-38.

[41]. Jatain, A. Chaudhary, S. Batra, P. Bhaskar, S. (2021) Rest web services: An elementary learning. Research Journal of Engineering and Technology, 12(3), 75-78.

[42]. Mor, P. Bhaskar, S. (2021) Enabling Technologies and Architecture for 5G-Enabled IoT. Blockchain for 5G-Enabled IoT: The new wave for Industrial Automation, 223-259.

[43]. Jatain, A. Chaudhary, S. Nagpal, P. Bhaskar, S. (2021) Cloud Storage Architecture: Issues, Challenges and Opportunities. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.

[44]. T Joy, D. Kaur, G. Chugh, A. Bhaskar, S. (2021) Computer Vision for Color Detection. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.
[45]. Jaglan, V. Bhaskar, S. (2021) Locking Paradigm in Hierarchical Structure Environment. Advances in Mechanical Engineering: Select Proceedings of CAMSE 2020, , 653-661.
[46]. Nanda, A. Gupta, S. Bhaskar, S. (2020) A Comprehensive Survey of Machine Learning in Scheduling of Transactions. 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(2020), 740-745.
[47]. Bhaskar S. Bhaskar, S. (2020) Study of locking protocols in database management for increasing concurrency. 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(2020), 556-560.
[48]. Sethi, N. Jaglan, V. Bhaskar, S. (2021) Multi-Operator based Saliency Detection. 2021 International Conference on Computational Performance Evaluation (ComPE), 507-512.
[49]. Swati, Jaglan, V. Bhaskar, S. (2021) A Novel Multi Granularity Locking Scheme Based On Concurrent Multi-Version Hierarchical Structure. Information Technology in Industry, 9(1), 932-947.
[50]. Sethi, N. Verma, J.K. Shrivastava, U. Bhaskar, S. (2021) Google Stock Movement: A Time Series Study Using LSTM Network. Multidisciplinary Functions of Blockchain Technology in AI and IoT Applications, 70-87.
[51]. Chaudhary, S. Jatain, A. Nagpal, P. Bhaskar, S. (2021) Design and Development of Gesture Based Gaming Console. A & V Publications, 12(2), 51-56.
[52]. Nagpal, P. Jatain, A. Chaudhary, S. Bhaskar, S. (2021) Motion detection of webcam using frame differencing method. A & V Publications, 12(2), 32-38.
[53]. Jatain, A. Chaudhary, S. Batra, P. Bhaskar, S. (2021) Rest web services: An elementary learning. Research Journal of Engineering and Technology, 12(3), 75-78.
[54]. Mor, P. Bhaskar, S. (2021) Enabling Technologies and Architecture for 5G-Enabled IoT. Blockchain for 5G-Enabled IoT: The new wave for Industrial Automation, 223-259.
[55]. Jatain, A. Chaudhary, S. Nagpal, P. Bhaskar, S. (2021) Cloud Storage Architecture: Issues, Challenges and Opportunities. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.
[56]. T Joy, D. Kaur, G. Chugh, A. Bhaskar, S. (2021) Computer Vision for Color Detection. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.