International Journal of Diabetes in Developing Countries (January–March 2025) 45(1):150–160
https://doi.org/10.1007/s13410-024-01334-4

ORIGINAL ARTICLE

Prediction of type II diabetes mellitus based on demographic features by the use of machine learning classification algorithms — a study across Assam, India

Partha Pratim Sarkar[1] · Snigdha Jyoti Das[2]

Received: 20 June 2023 / Accepted: 1 April 2024 / Published online: 17 April 2024

Abstract

Background The incidence of type II diabetes mellitus (T2DM) has quadruplicated in the recent decades and Prevention of T2DM cases is possible by changing lifestyle practices. The process of diagnosis of diabetes is a tedious one. The advent and advancement in (AI) techniques presents a probable solution to this critical problem.
Objective The study aims to assess the diverse attributes of the test sample population across Assam and enhance the early prediction of Type II Diabetes Mellitus by employing artificial neural networks.
Methods The aim of this study is to design a suitable AI model that prognosticates the likelihood of diabetes in individuals with maximum accuracy based on the levels of liver enzymes. This work also analyzes the effect of fast food intake, sleeping patterns, and consumption of alcohol on healthy controls and contemplates their susceptibility to contract T2DM.
Results The AI model accurately predicted T2DM likelihood and revealed significant links between unhealthy behaviors and increased T2DM risk among healthy individuals.
Conclusions The study underscores lifestyle modifications for T2DM prevention, highlighting AI's potential in diagnosis and the impact of unhealthy habits on T2DM susceptibility.

Keywords Artificial neural networks · Diabetes mellitus · Classification · Machine learning · Prediction

Introduction

In the past few decades, the number of people diagnosed with type II diabetes mellitus has multiplied about four times [1]. Type II diabetes mellitus is among the ninth major causes of death globally. T2DM affects approximately 90% of individuals globally [2], and the incidence rate in Asia is rising alarmingly. Projections indicate that within the next two decades, 70% of new T2DM cases will be in developing countries, predominantly in the age group of 45–64 years. Notably, seven of the top 10 countries with the highest diabetes burdens are lower-middle-income nations, including India, China, Russia, Brazil, Pakistan, Indonesia, and Bangladesh. India and China, with prevalence rates of 12.1% and 9.7%, respectively, stand out, as the primary hotspots [3]. As per Indian Council of Medical Research-India Diabetes (ICMR INDIAB), the prevalence of diabetes in India is 101 million. Type II diabetes mellitus is now emerging as a global epidemic, and the reasons for the increasing incidence of diabetes mellitus are numerous which includes ageing of the population, urbanization, economic development, consuming diets with improper nutrition, and sedentary lifestyles.

The onset of diabetes mellitus frequently occurs years before the diagnosis, even before it is actually diagnosed clinically. Approximately, 45.8% (or 174.8 million cases) of the total diabetes mellitus cases reported in adults were estimated to be undiagnosed. The people who are undiagnosed and untreated diabetes mellitus are at a greater risk of complications such as ketoacidosis and non-ketotic hyperosmolar than those who are receiving treatment.

✉ Partha Pratim Sarkar
Partha.sarkar@adtu.in

Snigdha Jyoti Das
dassnigdha725@gmail.com

[1] Faculty of Engineering, Assam Downtown University, Guwahati, Assam 781026, India

[2] Department of Molecular Biology and Biotechnology, Cotton University, Guwahati, Assam 781003, India

It has been reported that individuals diagnosed with type 2 diabetes have a higher occurrence of abnormal liver function test compared to individuals who are non-diabetic [4]. However, a chronic and mild elevation in the levels of transaminases (particularly ALT) often indicates underlying insulin resistance [5]. GGT (gamma-glutamyl transferase) is found to be elevated in patients with type 2 diabetes [6] that is another important and proposed non-specific marker of insulin resistance and type 2 diabetes [7]. It has a positive correlation with the intake of alcohol and smoking as well as coronary heart disease, BMI, systolic blood pressure, serum triglyceride, heart rate, and uric acid, when epidemiological data are taken into consideration. Studies also indicated that the level of ALT increased with increase intake of alcohol [8], thus acting as another secondary non-specific maker for insulin resistance and type 2 diabetes [9].

Snacking during late evening hours or having food at inappropriate hours has shown to increase daytime sleepiness, which correlates significantly with insulin resistance. Studies have also revealed that obstructive sleep apnea leads to elevated serum alanine aminotransferase (ALT) and aspartate aminotransferase (AST) levels [10, 11]. Food habits such as reduction in intake of fiber intake and increased intake of processed carbohydrates along with animal fats have been associated with development of obesity leading to excessive predisposition of type 2 diabetes [12].

Previous research has shown that type 2 diabetes mellitus (T2DM) is associated with lifestyle factors, yet there is a paucity of data regarding its prevalence among the Assamese population. With an increasing incidence of T2DM among the Indian population and considering the above background in knowledge, this study has been undertaken in the population of Assam, the state with anthropometric diversity. Also, recently with the advent and advancement of research in artificial intelligence, machine-learning algorithms have become a very popular tool in diagnosing diseases. Machine learning and data mining algorithms can process enormous amount of data and extract useful information pertaining to the study. Thus, the study has been designed to correlate the increased level of liver enzymes in diabetic patients as well as analyze parameters such as sleep pattern, food habits, and lifestyle pattern to prognosticate diabetes in healthy individuals using both probabilistic and machine-learning approaches. The performance of these methods is examined using various statistical metrics to achieve better accuracy.

Materials and methods

Study design

For the present cross-sectional and observational study, patients suffering from type II diabetes mellitus (n = 900) and healthy controls (n = 900), who were not diagnosed with T2DM, were enrolled with all clinical details from the Department of Medicine, Guwahati Medical College, Guwahati, under the supervision of registered medical practitioners with informed consent of the patients. Patients satisfying the inclusion criteria were enrolled in the study. The enrolled patients further underwent a designed health history, such as measurements of body mass index (BMI), sleeping duration, and fast food intake. All patients had liver function biomarkers, including alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), and gamma-glutamyl transferase (GGT). The data and clinical details used in the study were collected with patient's informed consent. The attributes of the dataset are presented in Table 1 below, based on the following inclusion and exclusion criteria:

Inclusion criteria

- Patients who were already diagnosed with type II diabetes mellitus were used for the study.
- Only those patients whose clinical data and informed consent was available were involved for the study.
- Patients in the age group 25–75 years were taken for the study.
- Voluntary individuals without type II diabetes mellitus were taken as healthy controls.

Exclusion criteria

- Patients below the age of 20 and above the age of 75 were excluded.
- Patients with any other known pathological infection were rejected.

Table 1 Details of the attributes employed in the dataset along with their Abbreviations

Sl. No	Dataset attributes	Abbreviations
1	Body mass index	BMI
2	Bilirubin(mg/dL)	BR
3	Total protein (g/dL)	TP
4	Albumin (g/dL)	A
5	Globulin (g/dL)	G
6	Alanine aminotransferase (IU/L)	AST
7	Aspartate aminotransferase (IU/L)	ALT
8	Alkaline phosphatase (U/L)	ALP
9	Gamma-glutamyl transferase (U/L)	GGTP
10	Fast food intake	FI
11	Alcoholic	A
12	Intake of non-veg	NV
13	Physical exercise	PE
14	Sleep duration (in hours)	S

- Patients with other serious systemic illnesses were also omitted.
- Patients without clinical data and informed consent were excluded from the study.

Sample size calculation According to the 2023 ICMR reports, the prevalence of type 2 diabetes mellitus in India stands at approximately 101 million. To conduct a comprehensive study within the framework of a larger ongoing research initiative, the sample size for both the disease population and controls was meticulously determined using Raosoft software. Employing a 95% confidence level and a 5% margin of error [4], the calculated sample size for the disease population alone was 1000 individuals. However, for the specific focus of the current study, the sample size was intentionally set at 900 for both the disease and control groups. Notably, controls were selected to be age and gender matched to ensure a robust and meaningful comparison [13]. This critical attention to demographic matching, coupled with a 95% significance level, contributes to the statistical strength of the study. Emphasizing the importance of precise data analysis, the chosen sample size of 900 for each cohort ensures the reliability and accuracy of the findings.

Artificial neural networks

Artificial intelligence (AI) techniques have become very popular and significantly prevalent in sectors ranging from business to healthcare. AI has the potential to transform the present state of healthcare dynamics with effective diagnosis, analysis, and interpretation of any concerned disease. Research has proven that AI is at par with human intelligence when it comes to the diagnosis of a disease. Algorithms have already started to outperform radiologists at spotting tumors and provide a bright spectrum for researchers to develop cohorts for costly clinic trials.

Artificial neural networks (ANNs) form the foundation of artificial intelligence and possess the ability to tackle intricate tasks that prove challenging for humans or conventional statistical methods. ANNs built upon the principles of the human brain, with neurons—the brain's functional components—serving as the fundamental units of an artificial neural network. These neurons operate in parallel and are organized into layers. The initial layer, functioning as the input layer, receives raw information, which is subsequently processed by successive layers. The final layers generate the output. Remarkably, neural networks can adapt to environmental changes, enabling enhanced learning within a system. A neuron *M* in mathematical terms can be understood with the help of the following equations:

$$u_m = \sum\nolimits_{i=1}^{n} w_{mi} x_i \quad (1)$$

And

$$\mathrm{y_m = \varphi(u_m + b_m)} \quad (2)$$

where $x_1, x_2, \ldots., x_i$ are the initial input signals; $w_{m1}, w_{m2}, \ldots.,$ w_{mi} are the signal strengths of neuron *M*, called as synaptic weights or simply weights; u_m combines the input signals to their corresponding weights linearly; the network's bias, denoted as b_m, plays a role in adjusting the final output y_m of neuron M, while the activation function $\varphi(.)$ determines how and when a neuron is initialized (Fig. 1a). The bias is a constant value that helps align the model with the provided data by modifying the output u_m. The activation function evaluates the weights associated with the neuron and adds the bias to the sum, thereby making decisions about neuron initialization. The activation function adds non-linearity into the output of the neuron. For detailed information on the working of ANNs, readers are referred to Haykins [14].

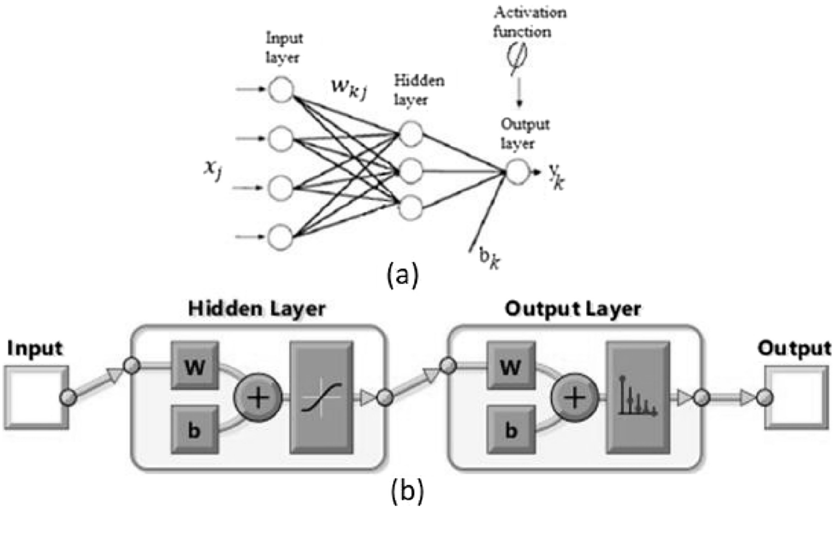


Fig. 1 **a** Functioning of a neural network and **b** a two-layer feed-forward neural network for pattern recognition trained with scaled conjugate gradient backpropagation algorithm

A two-layer feed-forward network (Fig. 1b), with sigmoid activation function in the hidden layer and softmax activation function, has been employed at the output layer. The network was trained with scaled conjugate gradient backpropagation algorithm. Neurons can classify vectors arbitrarily well, given enough neurons in its hidden layer. The given dataset set was divided into three parts, namely, training set (70%), testing set (15%), and validation set (15%). The best performance was achieved when the number of hidden neurons were set at 20.

Bayes probabilistic classifier

Bayes probabilistic classifier is a machine-learning model based on probability that serves as a classifier. It comes under supervised learning algorithms. This algorithm is based on the Bayes theorem. This classification method has a concept that delineates all features to be non-partisan and extraneous. This classifier that is based on conditional probability and therefore can be a very powerful tool for classification problems. The mathematical equation of Bayes theorem is given below:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P\left(\frac{A}{B}\right)}{P(B)} \tag{3}$$

where:

$P(A|B)$: probability of target class, $P(B|A)$: probability of predictor class, $P(A)$: probability of class A, and $P(B)$: prior probability of class B.

Support vector machines (SVM)

Support vector machines popularly known as SVMs also are supervised learning models. They are used for classification and regression analysis trained by associative learning algorithms. Vladimir N. Vapnik and Alexey Ya introduced the original form of the SVM algorithm in 1963. SVM is different from other algorithms for the fact that it uses a hyperplane that acts a decision-making boundary between different classes. SVMs have the intrinsic property to generate multiple hyperplanes for classification purposes such that every section consists of a particular class of data. The training of SVM is achieved on labelled data after which it can classify any unseen data depending on training prowess. SVMs are used for both classification and regression purposes. The classification in SVMs is achieved through the kernel trick. The kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data. Through the kernel trick, the data is transformed into another dimension so that it has a clear divide between different classes and then a hyperplane is set.

Results

The results segment composes of two sections: The former section focuses on the different models used in the study to diagnose type II diabetes in healthy individuals by analyzing various metabolic and physical traits, while the latter section focuses on the observational inference derived from the dataset.

Model results

The feed-forward network correlates the abovementioned parameters with greater accuracy in comparison to the other models. The performance of models evaluated with various statistical metrics is shown in Table 2 below. The accuracy metric assesses the effectiveness of algorithms in making predictions for individual instances. Cross-entropy is used to evaluate the performance of the network based on the target and output, as well as the synaptic weights and other parameters. Lower cross-entropy values indicate that the model has successfully generalized the classification problem. The same can be considered for the percentage error metric which is low during testing and validation. The receiver operating characteristic (ROC) curve for true positive vs false positive rate is also shown in Fig. 2. The ROC curve illustrates the diagnostic ability of our classifier as the discrimination threshold value varies. The area under curve (AUC) shows that our model performed efficiently during the test evaluation. The confusion matrix shown in Fig. 3 below shows the accuracy of the models in classifying parameters related to healthy individuals and individuals suffering from type II diabetes mellitus. ANN outperforms the other two models, which is evident from the confusion matrix shown below. It has the least percentage of false positive and false negative classifications. The percentage of misclassified parameters is significantly low for the test set. Overall misclassifications are observed to be at 8.1%. The recall metric measures the sensitivity of the classifier and again highlights the completeness of the ANN model in this task. From the performance of the models, it is inferred that ANN and Bayes outperform SVM in every aspect. The higher accuracy of ANN model is due to the fact it is more robust than the other two and is

Table 2 Performance of the employed models evaluated on various statistical metrics

Models	Accuracy	ROC	Cross-entropy	Percentage error	Recall
Bayes	0.78	0.81	3.12051e – 0	12.12	0.76
SVM	0.55	0.57	9.99377e – 0	39.32	0.59
ANN	0.83	0.87	1.12245e – 0	8.56	0.81

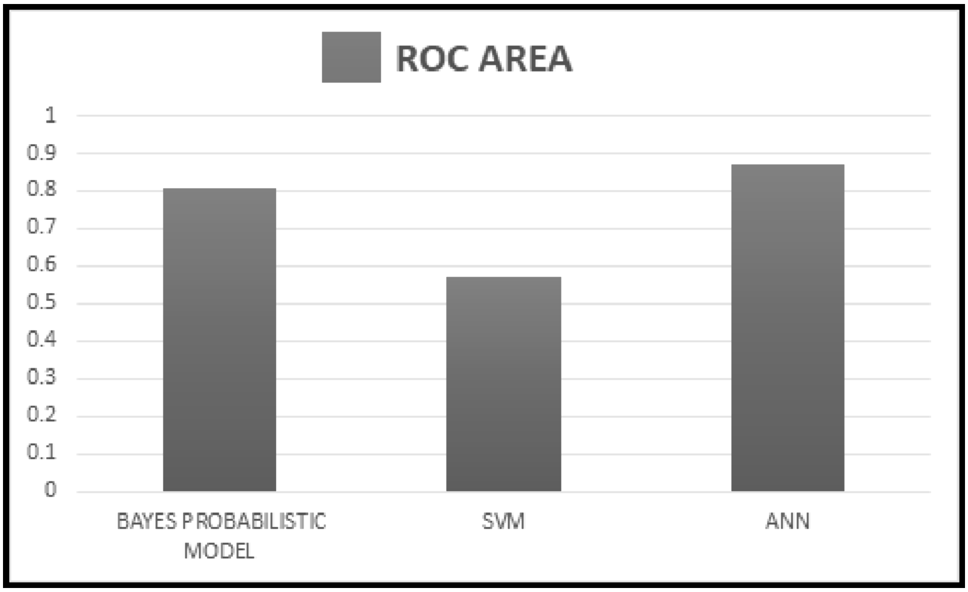
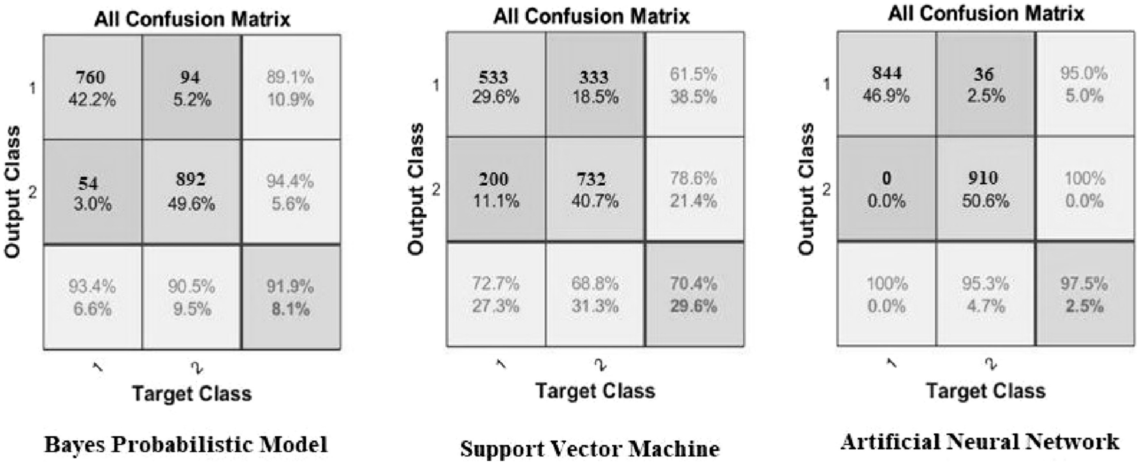


Fig. 2 Receiver operating characteristic (ROC) area of all the classifier models


All Confusion Matrix
Output Class
Target Class

Bayes Probabilistic Model

Output Class \ Target Class	1	2	
1	760 42.2%	94 5.2%	89.1% 10.9%
2	54 3.0%	892 49.6%	94.4% 5.6%
	93.4% 6.6%	90.5% 9.5%	91.9% 8.1%

Support Vector Machine

Output Class \ Target Class	1	2	
1	533 29.6%	333 18.5%	61.5% 38.5%
2	200 11.1%	732 40.7%	78.6% 21.4%
	72.7% 27.3%	68.8% 31.3%	70.4% 29.6%

Artificial Neural Network

Output Class \ Target Class	1	2	
1	844 46.9%	36 2.5%	95.0% 5.0%
2	0 0.0%	910 50.6%	100% 0.0%
	100% 0.0%	95.3% 4.7%	97.5% 2.5%

Fig. 3 Confusion matrix for the models, namely Bayes model, support vector machine, and artificial neural network

fault intolerant. The property of ANNs to create modifications within its system to change according to the system and subsequent learning comes in handy during the classification process.

Comparison of the level of liver enzymes in healthy controls and patients suffering from T2DM

This segment analyzes the correlation among the probable parameters that lead to the pathogenesis or increases the risk of type 2 diabetes mellitus in healthy individuals. The comparative analysis of the results of the level of liver enzymes between the two cohorts is presented in the upcoming sections.

ALP levels in healthy controls and patients suffering from type 2 diabetes mellitus

In previous studies that have been carried out, it is perceptible that the level of ALP in type II diabetic patients is elevated (ALP > 104 U/L) as compared to those who do not suffer from it [15]. However, in this study, it was observed that there was a significant increase in the level of ALP enzyme in the non-type II diabetic participants as well, evident from Fig. 4 given below. When the rate of incidence of patients suffering from type II diabetes was compared with the level of ALP, it was found that the number of individuals suffering from type 2 diabetes is highest, having ALP level of about 170 U/L. The number of type II diabetic individuals however remained less, though not significant, for the other

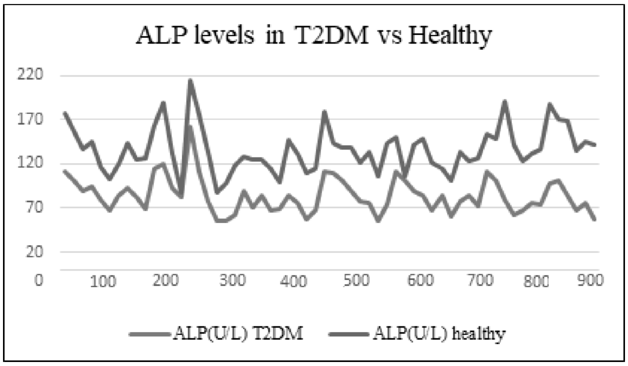


Fig. 4 Level of ALP in healthy controls and patients suffering from type II diabetes mellitus

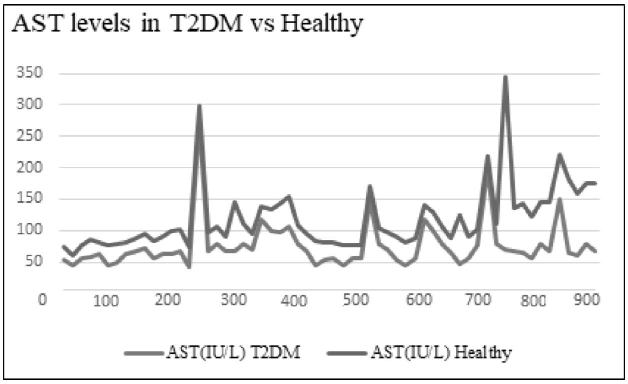


Fig. 5 Level of AST in individuals suffering from type 2 diabetes and normal healthy controls

levels of ALP. In the case of healthy individuals, ALP levels were around 120 U/L.

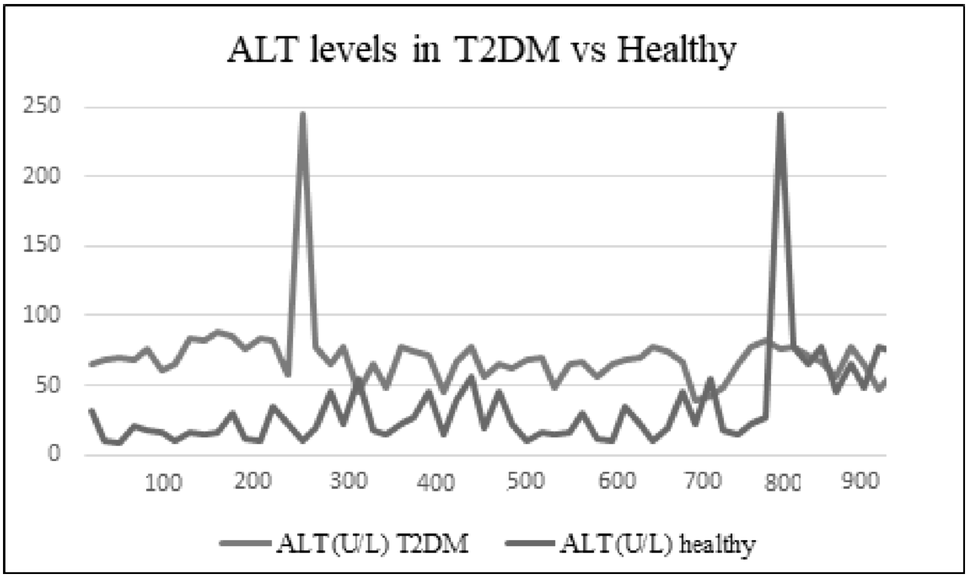
AST levels in healthy controls and patients suffering from type 2 diabetes mellitus

From Fig. 5, it can be observed that the level of AST is comparatively low in individuals suffering from type II diabetes. From the figure, it is evident that the highest incidence of individuals suffering from type II diabetes [16] is observed among those who have an AST level of about 300 U/L. This value coincided in non-type II diabetic individuals as well. Another striking observation that can be made from the figure is that the number of individuals having AST level of about 350 U/L was high in non-type II diabetic cohort compared to those suffering from type II diabetes. The number of individuals having other lower levels of AST was not significantly different in both the cohorts.

ALT levels in healthy controls and patients suffering from type 2 diabetes mellitus

In previous studies that have been conducted, it was found that a higher level of ALT (ALT > 56 U/L) [11] is an indicator of the pre-diabetic condition in individuals [17, 18]. In the study that has been carried out, the level of ALT was high in the individuals suffering from type II diabetes. From Fig. 6 shown below, it is observed that the number of type II diabetic individuals having ALT level approximately 250 U/L is highest. However, the level of ALT was found to be higher in non-type II diabetic individuals as well. The number of individuals who were not suffering from type II diabetes was highest having an ALT level of about250 U/L, which is nearly same as those individuals suffering from T2DM.

Fig. 6 Comparative analysis between the levels of ALT in individuals suffering from T2DM and non-T2DM individuals



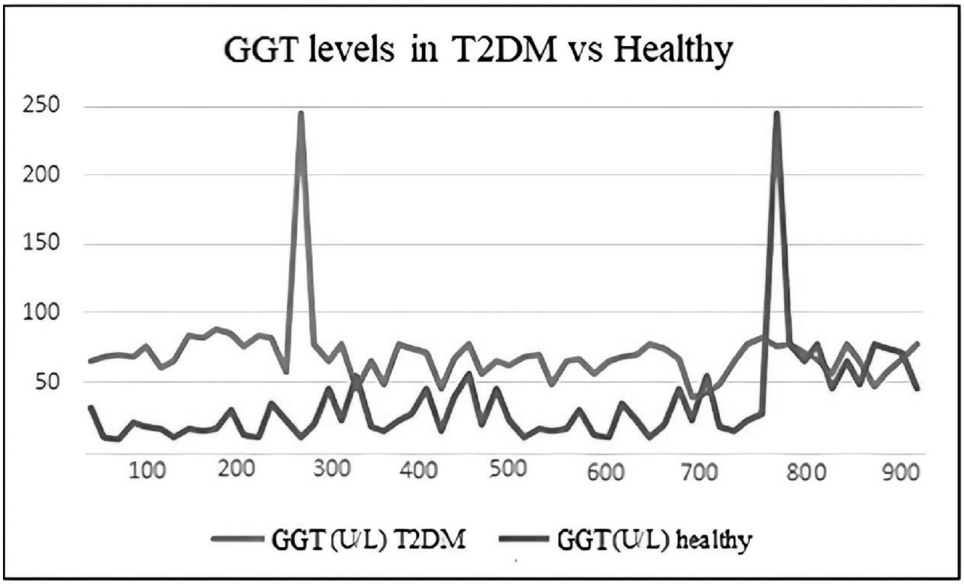


Fig. 7 Comparative analysis between the levels of GGT in individuals suffering from T2DM and non-T2DM individuals

GGT levels in healthy controls and patients suffering from type 2 diabetes mellitus

A similar previous research [19] has shown that the level of GGT increases in individuals suffering from type II diabetes. A similar trend in our study shows that the level of GGT is higher in individuals who were at a risk of developing diabetes or were in the pre-diabetic stage. It was observed that the level of GGT increased (GGT > 48 U/L) in individuals suffering from type II diabetes. Furthermore, from Fig. 7, we can infer that the number of individuals having type II diabetes show GGT levels around 750 U/L.

Effect of alcohol intake on the level of GGT

Previous studies have already shown that the level of GGT increases with increase in alcohol intake. The level of GGT has been shown to increase if the alcohol intake increases more than 20 mg/day. Increase in the level of GGT is indicative of the risk of individuals to develop type 2 diabetes mellitus. In the study that was carried out in non-type 2 diabetes individuals, it was found that the level of GGT in individuals increased positively with increase in alcohol intake. In Fig. 8, a correlation analysis was carried out between the level of GGT and alcohol intake in non-type 2 diabetic individuals which clearly indicates that the graphs coincided at

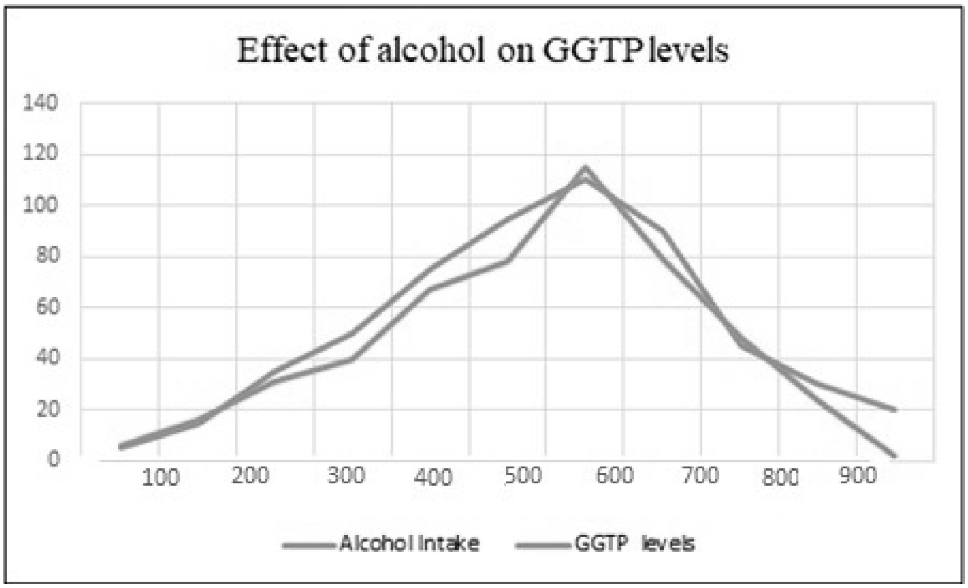


Fig. 8 Correlation analysis between alcohol intake and GGT level

a point where the levels of both alcohol intake and the level of GGT are highest, suggesting the point that the level of GGT increases with increase in alcohol intake.

Effect of sleeping duration on the level of liver enzymes

In a study carried out by Ioja [20], it was observed that the level of liver enzymes, particularly GGT and ALT [21], was found to be elevated in 42.3% of the study sample who had irregular sleeping pattern and less duration of sleep. In the demographic analysis that we carried out, it was found that 42.85% of the patients suffering from T2DM have irregular sleeping patterns as well as less duration of sleep (<4 h). This data was then correlated with the levels of liver enzymes which showed that they negatively correlated with proper sleep pattern. It was found that 7–8 h of proper sleep lead to a decrease in the level of liver enzymes.

From Table 3, it is understood that the percentage of increase in the various liver enzymes negatively correlates with the decreased duration of sleep, which is a trend observed in the non-type II diabetic individuals enrolled in the study. The level of the liver enzymes was found to be highest in individuals who had irregular sleeping patterns as well as slept for a duration ≤ 4 h. Out of the four mentioned liver enzymes, the level of ALT and GGT was found to increase at a maximum in individuals who had an average sleep duration of less than 4 h.

Effect of fast food intake on the level of ALT

Past research [22] has shown that the level of liver enzymes increases with increase in fast food and non-vegetarian food intake. Thus, an increase in level of ALT increases the risk of developing T2DM in healthy individuals. We have observed the level of ALT was higher in those individuals with high fast food intake. This is perceptible from Fig. 9 that shows a high correlation of 0.71 between the levels of ALT and fast food intake.

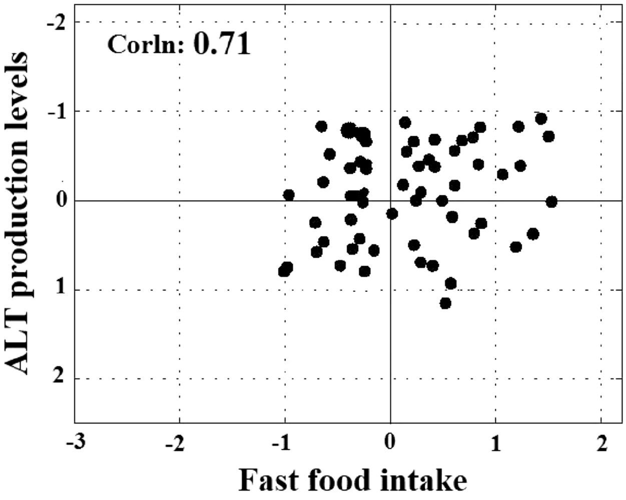


Fig. 9 Correlation between fast food intake and level of ALT in healthy individuals

Discussion

Diabetes mellitus and its complications are a major concern. In 2015, the International Diabetes Federation (IDF) estimated that in the age group of 20–79 years, one in every 11 adults (approx. 410 million) had diabetes mellitus all across the world. By 2040, this estimated to shoot up to 650 million, with the highest increases from the regions undergoing economic transitions from low-income to middle- or high-income levels. Escalation in the epidemic of diabetes mellitus is due to multiple reasons, which includes ageing of the population, urbanization and economic development, unhealthy food habits, and sedentary lifestyles. More than 90% of diabetes mellitus cases are type II diabetes mellitus (T2DM).

Table 3 Percentage increase in the levels of various liver enzymes with decrease in average sleeping duration in healthy individuals

Avg. duration of sleep (in hours)	Increase in AST levels (in %)	Increase in ALT levels (in %)	Increase in ALP levels (in %)	Increase in GGTP levels (in %)
4	28.37	29.72	26.31	29.85
4.5	24.32	21.62	20.12	22.10
5	16.21	18.91	17.10	16.35
5.5	13.51	12.16	11.06	10.35
6	10.81	9.45	9.75	8.21
6.5	5.40	2.70	3.41	2.05
7	2.70	2.71	2.45	2.01
7.5	1.35	1.40	1.37	1.25
8	1.33	1.39	1.25	1.05

However, the major real-world medical issue is the detection of diabetes at an early stage. Through this study, efforts were made to create models that serve the abovementioned purpose effectively. Three such models based on different algorithms were employed and tested on various statistical metrics. The model based on AI performed much better than its other counterparts did. The statistical model, Bayes model, also performed satisfactorily well given similar circumstances. With the advance in AI techniques, neural networks can be proposed as an effective diagnosing tool in detecting diabetes. In addition to the enzymatic parameters, the ANN model adequately mapped various other attributes such as alcohol intake, amount of sleep, and fast food intake with type II diabetes (correlational analysis). The higher accuracy metrics and true classifications between healthy and type II diabetic individuals confirm our notion.

Escalation in the epidemic of diabetes mellitus is due to multiple reasons, which includes ageing of the population, urbanization and economic development, unhealthy food habits, and sedentary lifestyles. More than 90% of diabetes mellitus cases are type 2 diabetes mellitus (T2DM). According to previous studies, the level of liver enzymes (ALP, ALT, and AST) is found to be elevated in individuals suffering from T2DM [1, 2]. This is in concordance with the present study where the subjects suffering from T2DM enrolled showed elevated levels of liver enzymes (ALP > 104 U/L, AST > 40 U/L, ALT > 56 U/L, GGT > 48 U/L). This is suggestive of the association of elevated levels of liver enzyme with the onset of T2DM and may be considered as a secondary prognostic marker. Approximately 52.06% of the healthy subjects enrolled in the study showed an elevation in the level of liver enzyme. The healthy subjects showed an approximate of 170–220 U/L in the level of ALP, 350 U/L of AST, 250 U/L of ALT, and 1450 U/L of GGT, respectively. The trend that has been observed in the healthy subjects is suggestive of the risk of development of pre-diabetic condition in them. These findings underscore the medical relevance of monitoring liver enzyme levels as potential indicators and prognostic markers for the onset and progression of T2DM.

In a previous study carried out by Pamidi et al. [22], it has been demonstrated that obstructive sleep apnea (OSA), a reversible sleep disorder, has become a novel risk factor, related to insulin resistance and glucose intolerance leading to the onset of pre-diabetes in approximately 20–67% population and T2DM in approximately 15–30% people. Several researches have also suggested that in patients suffering from T2DM, approximately 36–60% of them suffer from OSA than normal healthy people. In the present study, about 42.8% of the patients suffering from T2DM and 56.44% had irregular sleeping pattern with sleep duration of less than 4 h which is in line with previous researches [3]. These findings underscore the significance of recognizing the probable interplay between OSA and T2DM, as well as the impact of inadequate sleep duration, shedding light on potential avenues for intervention and preventive measures in the management of diabetes.

Previous researches have reported that the level of liver enzymes, particularly GGT and ALT, was elevated in 42.3% of the study sample who had irregular sleeping pattern and also less duration of sleep [5, 23]. When the level of liver enzyme and sleep pattern was analyzed in the diseased cohort in our study, it was found that majority of the patients suffering from T2DM had elevated levels of liver enzymes due to irregular sleeping pattern. In the healthy volunteers, it was observed that the level of liver enzyme was elevated in individuals having less than 4 h of sleep (ALT ~29.72%, AST ~28.37%, GGT ~29.85%, and ALP ~26.31%). It was also observed that with increase in sleep duration (≤8 h), the level of liver enzymes gradually decreased (ALT ~1.39%, AST ~1.33%, GGT ~1.05%, and ALP ~1.25%). This is suggestive of the fact that proper sleeping pattern is a modifiable risk factor associated with T2DM.

Previously carried out studies have shown that the level of GGT increases by twofold after heavy alcohol consumption [6]. This is in line with the observation in this study that indicated an elevated level of GGT in these individuals correlated positively with increase in alcohol intake. Interestingly, it has been concluded in earlier studies that elevation in the level of GGT is a secondary prognostic marker for predicting the onset T2DM or pre-diabetes in healthy individuals. Thus, the increased level of GGT can be used as prognostic marker in predicting the onset of T2DM in the current study population.

High intake of carbohydrates and fats, particularly from processed and red meats, contributes to insulin resistance and the development of type 2 diabetes mellitus (T2DM) [7]. Studies consistently indicate that diets low in fiber and high in glycemic content are associated with an elevated risk of T2DM [7]. The persistent consumption of processed meats and soft drinks further heightens the risk. In our study, a substantial 67.98% reported a high intake of fast food, indicating an increased risk of insulin resistance and T2DM onset. Notably, individuals with energy intakes exceeding 11.39% kcal/week from fast foods exhibited more than a twofold increase in ALT levels (≥12 U/L). In non-T2DM individuals, we observed a significant positive correlation ($p < 0.007$, $p < 0.05$) between fast food intake and elevated ALT levels. This suggests the pivotal role of diet as a modifiable risk factor in insulin resistance and T2DM development, emphasizing the importance of promoting diets rich in whole grains, fruits, vegetables, legumes, and nuts for diabetes prevention [7, 8].

The study revealed elevated liver enzyme levels, specifically ALT and GGT, in individuals with high alcohol and fast food intake, along with an irregular sleep pattern (less than 4 h). These findings suggest that apparently healthy individuals, not diagnosed with type 2 diabetes mellitus, are at a heightened risk of predisposition to the condition. Recognizing these risk factors is crucial for early intervention and diabetes prevention.

Conclusion

Through this study, we have created models both based on statistics and machine learning algorithms to help predict type II diabetes at an early age. Neural networks were observed to be performing better than Bayes and SVM model. In the future, such models can be used to diagnose or predict other similar diseases. However, the accuracy of machine learning models depends on the training and learning the task. These models can show better accuracy with a higher dataset. Therefore, for a better generalization capability, sufficient data has to be made available to these models. We also observed that the level of liver enzymes, particularly ALT and GGT, was elevated in individuals who have high alcohol and fast food intake as well as those having an irregular sleeping pattern and an average sleeping duration of less than 4 h. These traits are indicative of the fact that these healthy individuals, who have not yet been diagnosed with type II diabetes mellitus, are at a higher risk of predisposition towards it.

This study will help in establishing the facts that demographic factors, such as increased fast food intake, alcohol consumption, and irregular sleeping habits, play an indispensable role towards the pathogenesis of type II diabetes mellitus. A timely diagnosis and management of the abnormal liver enzyme levels, due to irregular and improper lifestyle patterns, may help to minimize the incessant rate of increase in the diabetic population.

Acknowledgments We sincerely acknowledge the Department of Medicine, Assam Medical College, Dibrugarh, Assam, India, for providing us the data sets pertaining to this study. The data was acquired under the supervision of registered medical practitioners with informed consent of the patients and will be produced if required/asked.

Data availability The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Declarations

Ethics approval Ethical approval from the Institutional Ethics Committee of Gauhati Medical College and Cotton University has been taken and will be provided if required.

Patient consent Required consent has been duly taken and will be provided if required.

IEC number for Gauhati Medical College and Hospital: MCI/190/2007/Pt-II/Oct.2022/44
IEC number for Cotton University: CU/ACA/ETHICS/2022/1

Conflict of interests The authors declare no competing interests.

References

1. Rafaqat S, Sattar A, Khalid A, Rafaqat S. Role of liver parameters in diabetes mellitus - a narrative review. Endocr Regul. 2023;57(1):200–20.
2. Li Y, Wang J, Han X, Hu H, Wang F, Yu C, Yuan J, Yao P, Li X, Yang K, Miao X, Wei S, Wang Y, Chen W, Liang Y, Zhang X, Guo H, Yang H, Wu T, He M. Serum alanine transaminase levels predict type 2 diabetes risk among a middle-aged and elderly Chinese population. Ann Hepatol. 2019;18:298–303.
3. Abdu Y, Naja S, Mohamed Ibrahim MI, Abdou M, Ahmed R, Elhag S, Saleh AO, Yassin M, Bougmiza I. Sleep quality among people with type 2 diabetes mellitus during COVID-19 pandemic: evidence from Qatar's National Diabetes Center. Diabetes Metab Syndr Obes: Targets Ther. 2023;16:2803–12.
4. Charan J, Biswas T. How to calculate sample size for different study designs in medical research? Indian J Psychol Med. 2013;35(2):121–6.
5. Yang J, Zhang K, Xi Z, et al. Short sleep duration and the risk of nonalcoholic fatty liver disease/metabolic associated fatty liver disease: a systematic review and meta-analysis. Sleep Breath. 2023;27:1985–96.
6. Fakhari S, Waszkiewicz N. Old and new biomarkers of alcohol abuse: narrative review. J Clin Med. 2023;12(6):2124.
7. Gu X, Drouin-Chartier J-P, Sacks FM, Hu FB, Rosner B, Willett WC. Red meat intake and risk of type 2 diabetes in a prospective cohort study of United States females and males. Am J Clin Nutr. 2023;118(6):1153–63.
8. Zhang Y, Meng Y, Wang J. Higher adherence to plant-based diet lowers type 2 diabetes risk among high and non-high cardiovascular risk populations: a cross-sectional study in Shanxi, China. Nutrients. 2023;15(3):786.
9. Mathur S, Mehta DK, Kapoor S, Yadav S. Liver function in type-2 diabetes mellitus patients. Int J Sci Stud. 2016;3:43–7.
10. Byrne TJ, Parish JM, Somers V, Aqel BA, Rakela J. Evidence for liver injury in the setting of obstructive sleep apnea. Ann Hepatol. 2012;11:228–31.
11. Prati D, Taioli E, Zanella A. Updated definitions of healthy ranges for serum alanine aminotransferase levels. Ann Intern Med. 2002;137:1–10.
12. Li J, Sun C, Liu S, Li Y. Dietary Protein Intake and Type 2 Diabetes Among Women and Men in Northeast China. Sci Rep. 2016;6:37604. https://doi.org/10.1038/srep37604
13. Das SJ, Dutta S, Tiwari D, Basumatary TK, Kashyap N, Kalita MP, Bose S, Bose PD. Prognostic value of myeloid differentiation primary response protein 88 in type II diabetes mellitus in non-obese NAFLD: a case-control study from Assam. Hum Gene. 2024;39:201246. https://doi.org/10.1016/j.humgen.2023.201246.
14. Haykin S. Neural networks: a comprehensive foundation. New Jersey: Prentice Hall PTR; 1998.
15. Gonem S, Wall A, De P. Prevalence of abnormal liver function tests in patients with diabetes mellitus. Endocrine Abstr. 2007;13:175.
16. Clark JM, Brancati FL, Diehl AM. The prevalence and etiology of elevated aminotransferase levels in the United States. Am J Gastroenterol. 2003;98:960–7.
17. Friedman LS, Dienstag JL, Watkins E. Evaluation of blood donors with elevated serum alanine aminotransferase levels. Ann Intern Med. 1987;107:137–44.
18. Erbey JR, Silberman C, Lydick E. Prevalence of abnormal serum alanine aminotransferase levels in obese patients and patients with type 2 diabetes. Am J Med. 2000;109:588–90.

19. Lebovitz H, Kreider M, Freed M. Evaluation of liver function in type 2 diabetic patients during clinical trials: evidence that rosiglitazone does not cause hepatic dysfunction. Diabetes Care. 2002;25:815–21.
20. Ioja S, Weir ID, Rennert NJ. Relationship between sleep disorders and the risk for developing type 2 diabetes mellitus. Postgrad Med. 2012;124:119–29.
21. Feng SZ, Tian JL, Zhang Q, Wang H, Sun N, et al. An experimental research on chronic intermittent hypoxia leading to liver injury. Sleep Breath. 2011;15:493–502.
22. Duffey KJ, Popkin BM. Adults with healthier dietary patterns have healthier beverage patterns. J Nutr. 2006;136:2901–7.
23. Pamidi S, Tasali E. Obstructive sleep apnea and type 2 diabetes: is there a link? Front Neurol. 2012;3:126. https://doi.org/10.3389/fneur.2012.00126

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.