Machine Learning-based Risk Prediction of Diabetes in Obese and Fatty Liver Patients: An Experimental Study

Raghda Essam Ali
Assistant Lecturer
Faculty of Computer Science
October University for
Modern Sciences and Arts (MSA)
Giza, Egypt
rrazek@msa.edu.eg, raghda.essam1@gmail.com

2023 Intelligent Methods, Systems, and Applications (IMSA) | 979-8-3503-3556-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IMSA58542.2023.10217744

***Abstract*—In recent years, diabetes has emerged as a highly significant global health concern, with a rapid increase in prevalence. It is projected that the number of people affected by diabetes will continue to rise, reaching 366 million by the year 2030, compared to 171 million in 2000. This study focuses on utilizing machine learning techniques to predict diabetes in patients with fatty liver disease. Four different machine learning classifiers, namely K-Nearest Neighbor, Fuzzy K-Nearest Neighbor, Support Vector Machine, and Artificial Neural Network, are employed to detect non-alcoholic fatty liver disease and predict diabetes. The analysis is performed on a genuine dataset obtained from Al-Kasr Al-Aini Hospital in Egypt, and the model's performance is evaluated using four-fold cross-validation. By comparing various machine learning algorithms, the study demonstrates that the Support Vector Machine classifier is particularly effective in identifying non-alcoholic fatty liver disease, achieving an accuracy rate of 95.8%. Furthermore, the Artificial Neural Network technique is employed to predict diabetic patients, yielding a model performance result of 86.6%. This research underscores the potential of machine learning techniques in healthcare for early prediction of chronic diseases and supporting decision-making processes.**

***Index Terms*—SVM, KNN, Fuzzy KNN, ANN, fatty liver, Obesity, Diabetes**

I. Introduction

Recent advancements in technology and data analysis have revolutionized the field of bio-medicine, enabling the acquisition and utilization of biomedical data to support medical decision-making processes. These advancements have also contributed to the development and advancement of computational software, which plays a pivotal role in interpreting and analyzing various types of medical data to aid in the decision-making process.

The primary objective of this paper is to leverage machine learning techniques to predict the occurrence of Diabetes Mellitus (DM) in obese patients by analyzing data patterns through classification methods.

In order to address the diabetes prediction problem, the study compares three different approaches: K-Nearest Neighbor (KNN), Fuzzy K-Nearest Neighbor (FKNN), and Support Vector Machine (SVM). Furthermore, the best-performing approach is combined with an artificial neural network (ANN) to enhance prediction accuracy.

The overarching goal of this paper is to employ machine learning techniques and analytics to empower healthcare professionals in timely identifying patients who are at risk of developing diabetes, particularly among individuals who are obese or suffer from non-alcoholic fatty liver disease to prevent the detrimental consequences of late diagnosis or inappropriate management of diabetes, which can lead to severe complications such as kidney failure, blindness, foot disorders, and even the necessity of amputation. It also raises the likelihood of heart disease, stroke, and premature death. In recent times, diabetes has emerged as one of the most critical health issues globally, with its prevalence expected to rise significantly in the coming years. In 2000 [1], 171 million people worldwide were diabetic, and this number is anticipated to reach 366 million by 2030.

This study is being conducted in Egypt, where the preva-

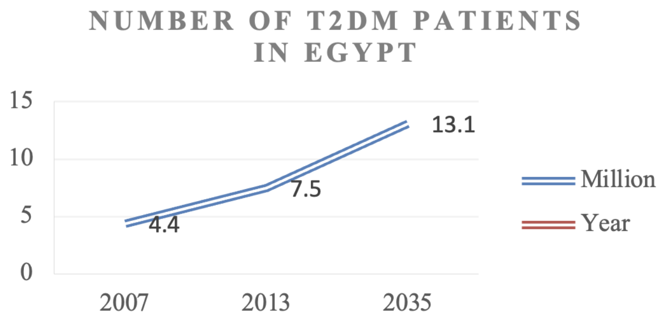

Fig. 1. the predicted development for Egypt's total number of T2DM patients by 2035.

979-8-3503-3556-9/23/$31.00 ©2023 IEEE

lence of type 2 diabetes mellitus (T2DM) among adults aged 20 to 79 years is currently at 16.8% according to the latest statistics of the international diabetes federation (IDF) in 2021 [2], with an annual diabetes mortality rate of 39.8 deaths per 100,000 populations. The Middle East and North Africa (MENA) region is expected to experience a significant increase in diabetic patients [3], [4], estimated to rise by 96% from 2013 to 2035. Egypt ranked as the 8th country in the world for the number of T2DM patients according to the IDF [2].

The IDF also reports that 7.5 million people in Egypt have diabetes and 2.2 million are pre-diabetic, but a significant portion of about 6.9 million people remains undiagnosed. The prevalence of diabetes in Egypt has risen rapidly over the years and is projected to increase to 13.1 million by 2035 as shown in Fig. 1. This highlights the importance of developing a technique to predict the risk of contracting this crucial disease, particularly in the aforementioned countries.

The traditional process of diagnosing potential diabetes is time-consuming and requires several diagnostic tests. This research aims to introduce an ML model to accurately predict the likelihood of diabetes before it occurs, using a two-step ML approach. Real-world data from Al-Kasr Al-Aini Medical Hospital in Cairo University is collected and analyzed to determine the best feature subset, assess the robustness of the machine learning algorithms to data uncertainties, and evaluate the predictive accuracy of the models.

Ultimately, this two-step ML approach can serve as a strong alternative for predicting and diagnosing diabetes and identifying relevant chronic diseases. Concerning this, the research discusses three primary research questions related to ML and its application in the medical field. Firstly, it examines how an ML approach can correlate between different diseases, such as liver disease, obesity, and diabetes. Secondly, it investigates whether applying different or multi-step ML approaches has a positive or negative impact on accuracy results. Finally, it explores whether reducing the number of features in ML affects its accuracy or not. By addressing these research questions, the research aims to enhance our understanding of how ML can be used to predict and prevent chronic diseases, ultimately improving patient outcomes and reducing the number of people affected by these illnesses.

The organization of this paper can be summarized as follows: Section II we will focus on the application of ML in the diagnosis and prediction of chronic diseases, with a particular emphasis on diabetes. The literature related to chronic disease and ML will be reviewed in section III. The proposed model, the ML techniques employed, and the methodology utilized are described in Section IV. Section V reports on the experiments and the results obtained from the proposed model. The conclusion is presented in Section VI, and Section VII outlines the future work that will be conducted in this field.

II. Background

Diabetes mellitus is a chronic disease characterized by hyperglycemia that can lead to a lot of complications. There are three types of diabetes; Type 1, Type 2, and Gestational Diabetes [5]. Type 2 diabetes is frequently associated with nonalcoholic fatty liver disease (NAFLD) [6], which can progress to nonalcoholic steatohepatitis due to obesity and insulin resistance. Obesity and diabetes are major public health concerns that can be prevented through lifestyle changes.

In healthcare research, machine learning has gained significant attention due to its ability to uncover insights by identifying hidden patterns using different data mining methods. Medical diagnosis is a suitable area for machine learning algorithms because it can identify patterns in extensive datasets. Machine learning and data mining techniques play a crucial role in using knowledge-based data in medical research, particularly for computer-aided diagnosis (CAD), as well as disease prevention and treatment.

Applying predictive methods to detect individuals with T2DM, NAFLD and obesity is crucial to minimize the severe consequences linked with this condition. Early identification is crucial for effective disease management and can promote interventions to prevent and manage diabetes mellitus and its related complications, especially in regions with limited resources.

III. Literature Review

The literature related to chronic disease and machine learning will be reviewed in this section, with a specific focus on studies related to diabetes prediction and diagnosis. By analyzing the findings of these studies, we aim to gain insights into the effectiveness of various ML algorithms and their potential in aiding healthcare professionals in accurate and early detection of diabetes. We will assess several predictive models and evaluate factors such as their techniques, performance, strengths, and weaknesses. A significant amount of research has been conducted to aid in disease detection, including the detection of diabetes [7]–[10], fatty liver [11], and the use of hybrid machine-learning techniques to predict one or more diseases [12] and have achieved high-accuracy results.

Alam, et al. [7], discuss the use of data mining techniques to predict diabetes at an early stage. The study selects significant attributes through principal component analysis and identifies a strong association of diabetes with body mass index (BMI) and glucose level using the Apriori method. The authors implement artificial neural networks (ANN), random forest (RF), and K-means clustering techniques to predict diabetes. The best accuracy of 75.7% was achieved by the ANN technique, which could potentially be used by medical professionals to improve treatment decisions.

Arumugam, et al., [8] introduced a study that evaluates the performance of three classification models: decision tree, Naïve Bayes, and Support Vector Machine (SVM). Precision, sensitivity, and specificity are measured for each model. The decision tree model achieves a precision of 78.37%, a sensitivity of 86.33%, and a specificity of 62.8%. The Naïve Bayes model achieves a precision of 79.89%, a sensitivity of 87.8%, and a specificity of 63.7%. The SVM model performs the best, achieving a precision of 84.89%, a sensitivity of 90%, and a specificity of 58.38%. The authors conclude that the design decision tree model, the method of Naïve Bayes, and the architecture of the SVM model result in the poorest predictive efficiency.

Wu, C.-C., et al. [11], made a comparison between different machine learning models, by including all patients at the New Taipei City Hospital having initial fatty liver between the period of 1st and 31st of December year 2009. Authors developed different classification models like Random Forest (RF), Naïve Bayes (NB), Artificial Neural Networks (ANN), and Logistic Regression (LR), in order to predict fatty liver disease (FLD). Results showed that the accuracy of NB, RF, LR, and ANN were 82.65%, 87.48%, 76.96%, and 81.85% respectively.

The study of Pradhan, N., et al. [9] proposes a system that uses publicly available data from diabetic patients to identify the causes, age groups, job styles, and eating habits associated with diabetes. The system employs Artificial Neural Networks to detect and recognize different types of diabetes. The study aims to predict the onset of diabetes in order to prevent associated health problems such as retinopathy, nephropathy, and cardiovascular conditions. The researchers tested their proposed model using the "Pima Indian Diabetes" dataset, which includes medical records of 768 patients and nine parameters for the development of diabetes. The proposed system achieved an accuracy of 85.09%, demonstrating its effectiveness in identifying diabetes.

Renganathan, V. [12], compared the performance of Artificial Neural Network (ANN) and Logistic Regression (LR) models in the biomedical field using an updated sample dataset from the Framingham Heart Study. The results demonstrated that the ANN model was more effective than the LR model in classifying the dependent variable. The ANN model achieved an accuracy of 84.4%, while the LR model had an accuracy of 82.9%.

The authors Sakib S, et al. in [10], utilized various ML algorithms on the PIMA Indian Diabetes Dataset, XGBoost was found to be the most accurate with an 80.73% accuracy rate, followed by SVM with a rate of 80.21%. The study aims to assist doctors and clinicians in early detection of diabetes using ML techniques.

Thus, from the literature it is clear that there are various ML algorithms used in the medical field for disease detection, and many studies have used a combination of classifiers to predict multiple diseases, achieving high accuracy results. Other studies have compared the performance of different ML techniques for predicting chronic diseases and achieved high accuracy results.

Considerings the relevant works and after showing the different ML classifiers performance, this research aims to evaluate the performance of four ML classifiers (ANN, SVM, KNN, and FKNN) and apply a two-step model to achieve the best accuracy result. It also aims to examine the consequences of using machine learning techniques in disease prediction and whether it can assist physicians in decision-making and avoiding the probability of disease occurrence.

IV. Proposed Model

This section addresses the proposed model. It also describes the preprocessing techniques that were carried out on the patient data. The proposed model is divided into two stages. The initial stage is focused on identifying individuals with nonalcoholic fatty liver disease (NAFLD), while the second stage is concerned with predicting the likelihood of patients developing diabetes. The model employs different ML approaches to enhance the performance of the classifier and improve its accuracy.

A. Data Collection

The dataset used in this study was obtained from Al-Kasr Al-Aini Hospital, Faculty of Medicine, Cairo University. The dataset is in the form of an Excel file (.xlsx) and contains a total of 260 records. Out of these records, 108 are classified as true positive and 152 as true negative. The dataset includes 32 features for analysis which are; age, gender, schistosomiasis (Shisto), smoking, alcohol consumption, oral contraceptive pill (OCP), steroids, family history of diabetes, history of hypertension, height, weight, body mass index (BMI), waist circumference (WC), liver disease, hemoglobin test (HGB), primed lymphocyte test (PLT), white blood cells (WBCs), basic Insulation level (BIL), albumin level in blood (ALB), aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), protein C test (PC), international normalized ratio (INR), total cholesterol, low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides test (TGs), fasting blood sugar (FBs), hemoglobin A1c (HBA1C) and spleen size. The dataset was divided into two steps according to the medical aspects the physicians referred that liver disease can indicate diabetes disease, then we can predict diabetes disease after detecting the fatty liver disease. Thus, the first set after applying feature selection and extraction [6] is consists of 9 selected features which are; age, gender, shisto, ALT, AST, ALP, GGT, INR, TGs and the desired class of the NAFLD. The second set consists of 7 selected features which are; the output of the NAFLD as an input feature to the second phase in addition to six more features which are height, weight, or BMI which is calculated from the patient's weight in kilograms over height in meter2 (kg/m2) [13], WC, FBS, history of hypertension and family history of diabetes.

B. Model Formulation

The suggested model showed in Fig.2. started from a preprocessing step of filtering data then estimating the missing values, standardizing data, normalizing data after that handling the imbalanced data then verifying data to finally be ready for feature selection and extraction. Then, the output of the model comprises two phases, the first phase was comparing three different algorithms; K-Nearest Neighbor, Fuzzy K-Nearest Neighbor, and Support Vector Machine algorithm to find the best accuracy result from each of which in order to detect patients with nonalcoholic fatty liver disease (NAFLD). The learning algorithms were applied to 9 selected features and

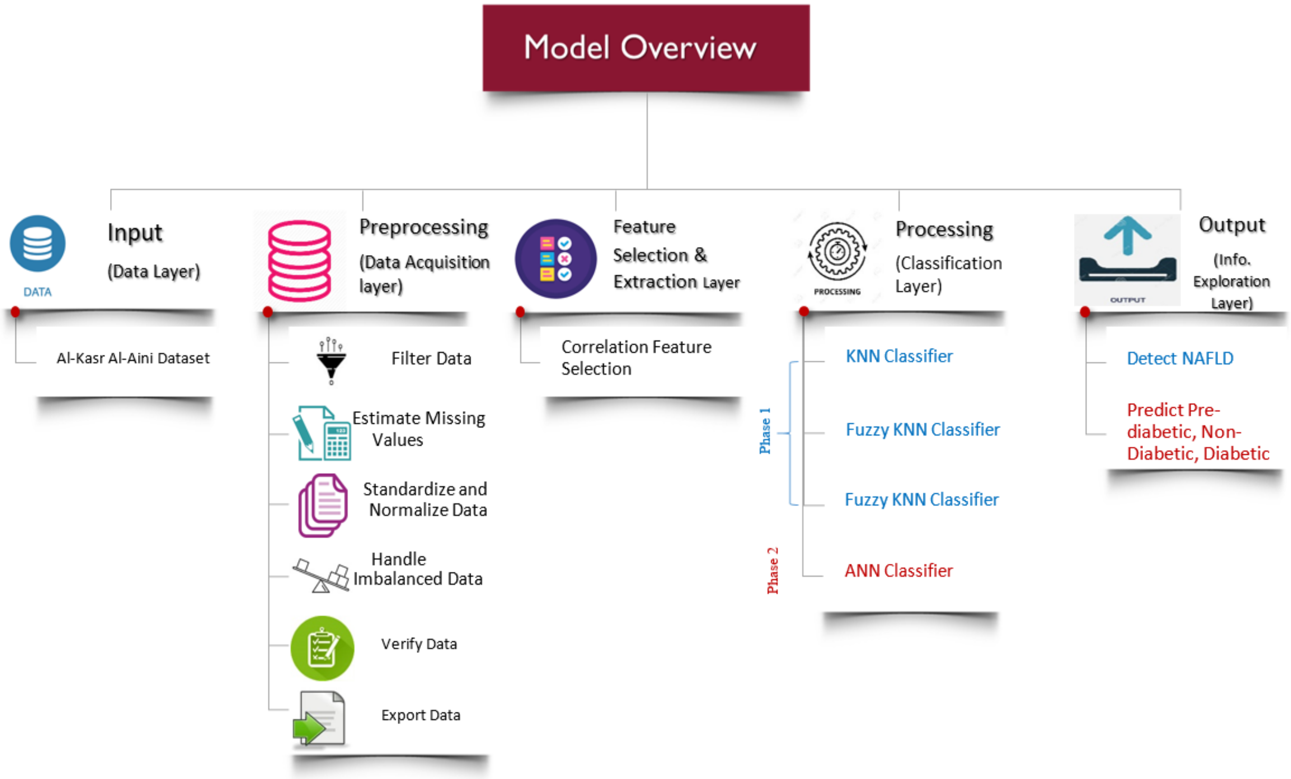

Fig. 2. Model Formulation

a number of patients with NAFLD were detected. Patients with other reasons for liver illness (alcohol, medication, etc.) were excluded to give off the results to be either 0: does not have non-alcoholic fatty liver disease or 1: affected with non-alcoholic fatty liver disease [6].

The second phase was applying a back propagation neural network that takes the dataset of NAFLD patients as input in phase 2 in addition to another 6 selected features. Then start the training process of the artificial neural network algorithm with distinct topologies and a range of epochs to achieve weights that give the optimum outcomes to categorize patients into three different classes; pre-diabetic, diabetic, or not-diabetic patients as shown in Fig.2.

V. Model Application & Results

In this section, we will discuss the accuracy of the different machine learning algorithms over the explained dataset in section IV.A. In the healthcare sector, it is crucial to predict diseases early to prevent their occurrence. DM is a particularly alarming disease that is on the rise, and ML techniques are being used to predict its likelihood of occurrence. Several algorithms, including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), and Fuzzy K-Nearest Neighbor, are commonly used for this purpose. SVM is a supervised learning model that sorts data into two categories for analysis, while ANN uses a network of functions to translate input data into the desired output. KNN is an algorithm that applies classification without prior knowledge of data, and Fuzzy K-Nearest Neighbor assigns class membership to each sample based on its distance from its nearest neighboring k-element in the training samples. FKNN is a popular choice among classifiers due to its simplicity and the certainty of its classification decisions. As mentioned in section IV.B, there are two model phases; the first one classifies the patient according to liver infection, using the learning algorithms on 9 selected features. The second classification used an ANN technique that takes the output from phase 1 as an input in phase 2, in addition to 6 additional selected features. The training process of the artificial neural network algorithm then categorizes patients into three different classes; pre-diabetic, diabetic, or non-diabetic patients.

A. ***Phase 1 Experiment:*** *Detecting Non-Alcoholic Fatty Liver Disease (NAFLD)*

In this phase, various machine learning algorithms (e.g., K-Nearest Neighbor, Fuzzy K-Nearest Neighbor, and Support Vector Machine) are applied and their results are compared in order to determine the best accuracy results. In the beginning, different values and functions were tested to find the best K-value for the KNN algorithm. As shown in Table I, the KNN classifier is obviously the most accurate with 91.92% accuracy if applied on 2-segment datasets of K-value with the correlation function. It decreases gradually at K=4, K=7, and K=10 to 85.00%, 73.67% , and 76.72% respectively.

TABLE I
KNN Algorithm Results on 9 Selected Features

Function	**K-Nearest Neighbor (KNN) Algorithm**			
	K=2	K=4	K=7	K=10
Eucledien	91.15%	83.84%	69.00%	71.90%
Cityblock	90.76%	82.30%	68.46%	73.07%
Cosine	91.53%	85.76%	75.38%	78.46%
Correlation	**91.92%**	85.00%	73.67%	76.72%

TABLE II
FKNN Algorithm Results on 9 Selected Features

Fuzzification parameter= 0.3	**K-Nearest Neighbor (KNN) Algorithm**			
	K=2	K=4	K=7	K=10
Eucledien	**91.15%**	84.23%	79.23%	80.00%

Then applying the Fuzzy KNN classifier requires a fuzzy parameter and a K-value. After experimenting with different values to determine the right value of K and fuzzy parameters, we applied 2 segments of K-value and 0.3 fuzzy parameters. This showed in Table II an accuracy of 91.15%.

It then declines at K=4, K=7, and K=10 to 84.23%, 79.23% and 80.00% respectively. The support Vector Machine algorithm performed the best accuracy result of 95.00% as shown in Table III. Fig. 3 shows a clear comparison between all three ML algorithms and shows the accuracy results in changes.

TABLE III
SVM Algorithm Results on 9 Selected Features

Results	**Support Vector Machine (SVM) Algorithm**			
	RBF	Gaussian	Linear	Polynomial
Accuracy	**95.00%**	**95.00%**	73.46%	89.23%

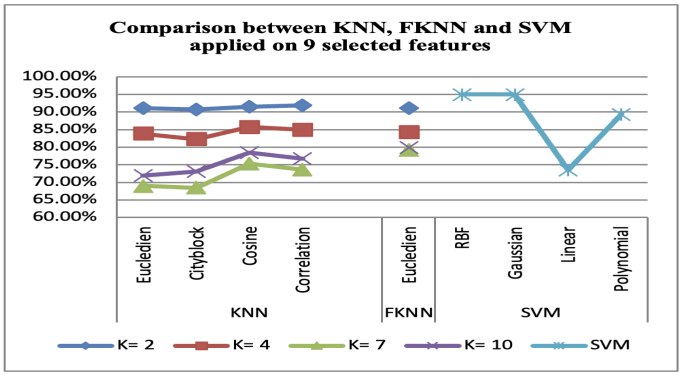

Fig. 3. Comparison between ML Techniques on 9 Selected Features

After applying the previous ML algorithms on the 9 selected features, it is clear that the accuracy result of the SVM algorithm using the RBF function shows the highest accuracy result compared to the other algorithms as shown in Fig.3. However, this rise up a question for us, what if we eliminate some of the selected features, will this affect the precision positively or negatively? This pushes us to re-examine the proposed model to check whether the precision will be increased or decreased. Thus we applied feature elimination using the correlation function which involves identifying the degree of correlation between the features in the dataset and removing highly correlated features that may cause multicollinearity issues in the model.

This can be achieved by calculating the correlation coefficient between each pair of features and removing one of the features of the correlation coefficient that is above a specified threshold as shown in Table IV.

The remaining features are then used for further analysis and model building as shown in Table V. As a result, we utilized the correlation function to determine the statistical correlation among the arbitrary features. Table IV indicates that all the chosen features display a strong correlation. Nonetheless, we will begin by removing the less correlated features, namely Gender, ALT, and TGs, as shown in Table V. Then, we will re-evaluate the proposed model by employing three machine learning algorithms and observing any changes in the accuracy results.

TABLE IV
Features Degree of Correlation

Feature	**Age**	**Gender**	**Shisto**	**AST**	**ALP**	**ALT**	**GGT**	**INR**	**TGs**
Age	1.00	-0.05	-0.01	0.09	-0.02	0.03	-0.08	-0.06	-0.01
Gender	-0.05	1.00	-0.10	-0.25	-0.31	-0.08	-0.08	-0.02	-0.13
Shisto.	-0.01	-0.10	1.00	-0.05	0.03	-0.06	-0.05	-0.06	0.04
AST	0.09	-0.25	-0.05	1.00	0.77	0.27	0.38	0.17	0.30
ALP	-0.02	-0.31	0.03	0.77	1.00	0.25	0.38	0.12	0.30
ALT	0.03	-0.08	-0.06	0.27	0.25	1.00	0.64	0.55	0.18
GGT	-0.08	-0.08	-0.05	0.38	0.38	0.64	1.00	0.45	0.40
INR	-0.06	-0.02	-0.06	0.17	0.12	0.55	0.45	1.00	0.30
TGs	-0.01	-0.13	0.04	0.30	0.30	0.18	0.40	0.30	1.00

TABLE V
Features Correlation Coefficient

Age	**Gender**	**Shisto**	**AST**	**ALP**	**ALT**	**GGT**	**INR**	**TGs**
-0.13	**0.25**	0.22	-0.22	-0.11	**-0.24**	-0.03	0.09	**-0.25**

As shown in Table VI, the KNN classifier is obviously the most accurate with 91.15% accuracy, if applied on 2-segment datasets of K-value with the correlation function. Then it decreases gradually at K=4, K=7, and K=10 to 83.46%, 71.54%, and 75.77% respectively.

TABLE VI
KNN Algorithm Results on 6 Selected Features

Function	**K-Nearest Neighbor (KNN) Algorithm**			
	K= 2	**K= 4**	**K= 7**	**K= 10**
Eucledien	88.08%	81.15%	69.23%	73.46%
Cityblock	90.77%	81.54%	70.77%	74.62%
Cosine	91.15%	81.92%	69.62%	72.69%
Correlation	**91.15%**	83.46%	71.54%	75.77%

After experimenting with different values to determine the right value of K and fuzzy parameters, we applied 2 segments of K-value and 0.3 fuzzy parameters, this showed in Table VII with an accuracy of 88.08%. It then declines at K=4, K=7, and K=10 to 82.31%, 76.15%, and 76.54% respectively.

TABLE VII
FKNN ALGORITHM RESULTS ON 6 SELECTED FEATURES

Fuzzification parameter= 0.3	Fuzzy K-Nearest Neighbor (KNN) Algorithm			
	K=2	K=4	K=7	K=10
Eucledien	88.08%	82.31%	76.15%	76.54%

The support Vector Machine classifier performed the best accuracy result of 95.76% as shown in Table VIII.

TABLE VIII
SVM ALGORITHM RESULTS ON 6 SELECTED FEATURES

Results	Support Vector Machine (SVM) Algorithm			
	RBF	Gaussian	Linear	Polynomial
Accuracy	95.76%	95.76%	67.69%	87.30%

Thus, When applying the previous ML algorithms on the 6 selected, the accuracy result of SVM algorithm using the RBF function shows again the highest accuracy result compared to the other algorithms as shown in Fig. 4. Moreover, it increased again by 0.76% than that resulted when applied the same algorithm on 9 features in Table III, while it increased only by 0.38% when applied to 6 features Table VIII. Accordingly, this slight enhancement shows that the three features; gender, TGs, and ALT are not strongly correlated.

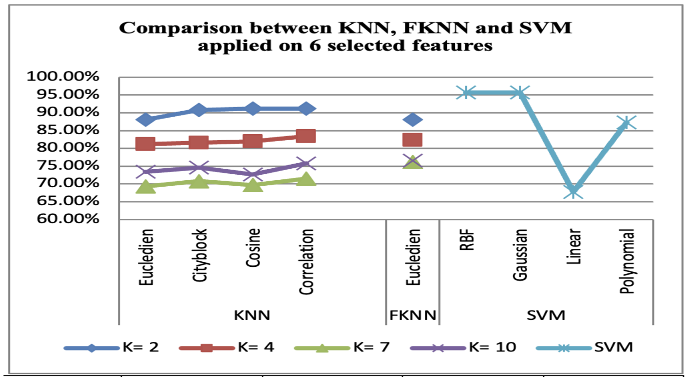

Fig. 4. Comparison between ML Techniques on 6 Selected Features

B. Phase 1 Results:

In this section, we will analyze all the results to find out which ML algorithm performs better. In section V.A, It is clearly noticeable that the SVM algorithm achieves the best accuracy result when applied on the 6 selected features; age, shisto., AST, ALP, GGT and INR that reached 95.76% compared to 95.00% that resulted when applied on the 9 selected features; age, gender, AST, ALP, ALT, GGT, INR and TGs.

On one hand, as shown in Table IX, we found that the KNN algorithm with k-value equals 2 and correlation function shows the best accuracy result when applied on 9 selected features by 91.92%. However, the Fuzzy KNN algorithm with k-value equals 2 and fuzzy parameter equals 0.3 and correlation function accuracy decreased by 0.77% to reach 91.15% when applied on the same selected features.

TABLE IX
ML TECHNIQUES ACCURACY RESULTS VS. ALL SELECTED FEATURES

KNN vs. FKNN vs. SVM Accuracy Results			
Algorithms	KNN	FKNN	SVM
Selected Features	Correlation Function at K=2	Fuzzy Parameter= 0.3 & K= 2	Gaussian Kernel Function
9	91.92%	91.15%	95.00%
7	90.76%	89.23%	95.38%
6	91.15%	88.08%	95.76%
5	90.77%	88.08%	94.23%
4	89.23%	90.23%	91.92%

On the other hand, we found that the SVM algorithm with Gaussian and RBF functions performs better accuracy result when applied on the same selected features and shows an increment by 3.08% than that of the KNN algorithm and by 3.85% compared to the Fuzzy KNN algorithm result. Although it shows the best accuracy when applied on 6 selected features only to increase the result expectations a little bit by 0.76% to achieve the highest precision of 95.76% as shown in Fig. 5.

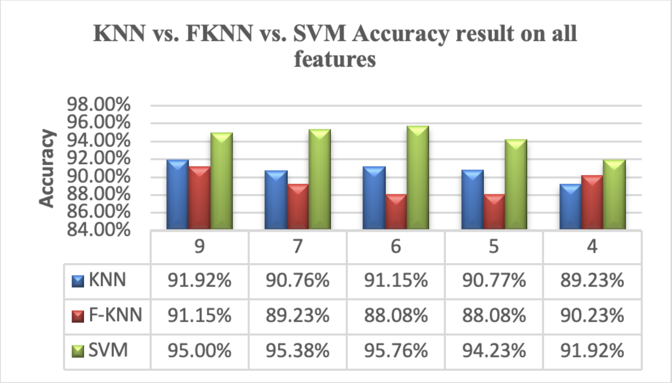

Fig. 5. KNN vs. FKNN vs. SVM Accuracy Results on all selected features

C. Phase 2 Experiment & Results: *Predicting Type-2 Diabetes Mellitus Disease (T2DM)*

In this phase The ANN classifier is used to model the diabetes chronic disease. The second phase was developing a back propagation neural network that takes the output from phase 1 as an input in phase 2 in addition to another 5 selected features which are; BMI, WC, FBS, History of hypertension and Family History of Diabetes [6], then start the training process of the artificial neural network algorithm with distinct topologies and range of epochs to achieve weights that give the optimum outcomes to categorize patients to three different classes; pre-diabetic, diabetic or not-diabetic patients.

The applied ANN classifier has i configuration, where i=7 (the number of features to the model input), hidden layer h where h is the number of neurons in the hidden layer, we applied the ANN using two hidden layers h1 & h2, where h1=8 and h2=7, and o is the number of outputs that is equal one. The best performance was achieved using the primary dataset with overall 7 features.

Then, the classifier is trained as showed in Table X, and demonstrates that the optimal accuracy results achieved in 50 iterations with 7 input features, 8 and 7 nodes in the hidden layer1 and hidden layer 2 respectively and 1 node in the output layer was 86.6% accuracy.

In the experiments, after analyzing the results, it can be concluded that the use of the two-step model that combines the SVM and ANN in one system is clearly preferable to using each classifier individually. As per the empirical study, when applying the ANN classifier alone on the same dataset, the accuracy result performed was 78.07% compared to 86.56% which was reached when applying the suggested model.

TABLE X
ANN Accuracy Results

Layer		**Artificial Neural Network**			
		Accuracy Results			
Input Features		7	7	7	7
Hidden 1		**3**	**5**	**7**	**8**
Hidden 2		**3**	**5**	**5**	**7**
Output		**1**	**1**	**1**	**1**
Iterations	**50**	81.53%	82.69%	83.84%	**86.56%**
	100	82.30%	82.69%	84.61%	85.38%
	150	83.07%	—	84.61%	85.38%
	200	83.07%	—	—	—

This proves that the two-steps model stated, improves the accuracy result by almost 8.5%.

Primarily, because SVM classifier is more efficient with binary class problems and very sensitive to the dimensionality of the feature vectors. In addition to the ANN algorithm which is supposed to be a very flexible classifier. This combination leads to a powerful technique for classification problems.

VI. Conclusion

This research paper aims to address the increased risk of severe health outcomes, such as cirrhosis and mortality, for patients with non-alcoholic fatty liver disease (NAFLD) and diabetes mellitus. The study proposes a two-phase model to predict chronic diabetes mellitus. In the first phase, three different algorithms (K-Nearest Neighbor, Fuzzy K-Nearest Neighbor, and Support Vector Machine) were compared to detect patients with NAFLD. The algorithms were able to accurately detect patients with NAFLD and exclude those with other liver diseases.

In the second phase, the study developed a back-propagation neural network to classify patients into three categories: pre-diabetic, diabetic, or not diabetic. The neural network was trained with selected features and the output from the first phase, which used the Support Vector Machine algorithm to detect NAFLD. The proposed model employs two machine learning techniques, the Support Vector Machine algorithm and the Artificial Neural Network algorithm. The obtained accuracy results showed that the model achieved 95.76% accuracy in detecting NAFLD in phase 1 and 86.56% accuracy in predicting diabetes in phase 2, using 7 input features, 8 and 7 nodes in the hidden layers 1 and 2, respectively, and 1 node in the output layer.

The study suggests that using only 6 selected features can minimize the effort, time, and cost of lab tests for patients. Patients with NAFLD and obesity are at higher risk of developing diabetes and other severe illnesses. The proposed model can help physicians in early disease prediction and reduce the number of diabetic patients worldwide. The model recommends suitable machine learning techniques for early diagnosis and appropriate treatment.

VII. Future Work

The study suggests that more experiments are needed in the future to improve the performance of the model. This includes creating a more balanced dataset by involving more data, as well as applying deep learning neural networks on a larger dataset for better accuracy. Additionally, more research is required to overcome challenges in medical research and to advance machine learning technology beyond disease prediction. The study also proposes expanding the work to aid in predicting not only diabetes but other chronic medical diseases.

References

[1] W. Organization, "Definition, diagnosis and classification of diabetes mellitus and its complications: report of a who consultation. part 1, diagnosis and classification of diabetes mellitus," *World Health Organization*, 1999.

[2] I. D. Federation, *IDF diabetes atlas*. Brussels: International Diabetes Federation, 2013.

[3] R. Hegazi and et al., "Epidemiology of and risk factors for type 2 diabetes in egypt," *Annals of global health*, vol. 81, no. 6, pp. 814–820, 2015.

[4] W. H. Organization, "Use of glycated haemoglobin (hba1c) in diagnosis of diabetes mellitus: abbreviated report of a who consultation," tech. rep., World Health Organization, Geneva, 2011.

[5] S. Padhi, A. K. Nayak, and A. Behera, "Type ii diabetes mellitus: a review on recent drug based therapeutics," *Biomedicine & Pharmacotherapy*, vol. 131, p. 110708, 2020.

[6] R. E. Ali and et al., "Prediction of potential-diabetic obese-patients using machine learning techniques," 2019.

[7] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, and Z. Abbas, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019.

[8] S. S. Arumugam, V. Kuppan, V. Chakravarthi, and K. Palaniappan, "An accurate diagnosis of diabetes using data mining," in *AIP Conference Proceedings*, vol. 2405, p. 020017, AIP Publishing LLC, 2022.

[9] N. Pradhan, S. Sinha, and J. P. Tripathy, "Diabetes prediction using artificial neural network," in *Deep Learning Techniques for Biomedical and Health Informatics*, pp. 327–339, Elsevier, 2020.

[10] S. Sakib, N. Yasmin, I. K. Tasawar, A. Aziz, M. A. A. Siddique, and M. M. R. Khan, "Performance analysis of machine learning approaches in diabetes prediction," in *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 1–6, IEEE, 2021.

[11] C.-C. Wu, Y.-C. Hung, H.-F. Yang, D.-Y. Wu, K.-H. Wu, K.-C. Chang, and P.-C. Chen, "Prediction of fatty liver disease using machine learning algorithms," *Computer methods and programs in biomedicine*, vol. 170, pp. 23–29, 2019.

[12] V. Renganathan, "Overview of artificial neural network models in the biomedical domain," *Bratislavske lekarske listy*, vol. 120, no. 7, p. 536, 2019.

[13] A. Y. Soeroto, N. N. Soetedjo, A. K. Purwiga, P. Santoso, H. Suryadinata, S. Setiati, and A. W. Sudoyo, "Effect of increased bmi and obesity on the outcome of covid-19 adult patients: A systematic review and meta-analysis," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 6, pp. 1897–1904, 2020.