International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 01 | Jan 2023 www.irjet.net p-ISSN: 2395-0072

Diagnosis of Diabetes Mellitus Using Machine Learning Techniques

Mr.P.Bhanu Chand1, M.Lalitha Kavya2, G.SAI SUSHMITHA3, M. KHYATHI SRI4,M. INDIRA5

[1]Assistant Professor Department of Information Technology, KKR & KSR Institute Of Technology And Sciences (A), Guntur, India
[2,3,4,5]Undergraduate Students ,Department of Information Technology , KKR & KSR Institute Of Technology And Sciences (A) ,Guntur, India

---

Abstract - Recently, many terrible diseases have affected human health. Many diseases are spreading and causing serious damage to mankind. Advances in technology have proven that most diseases can be cured in the age of medicine, but some diseases can only be prevented and cannot be cured, one of which is diabetes. This article reports on a medical case examining the electronic medical records of diabetics from various sources. Analysis was performed using two data mining classification algorithms: Random Forest and Support Vector Machine. The performance of the SVM algorithm is analyzed for different cores available. The best kernel is selected and used for prediction. Random forests are similar to bootstrap algorithms using decision tree models (CART). The purpose of the analysis is to predict diabetes using medical records and compare the accuracy of these two algorithms to find the best diabetes prediction algorithm.

Key Words*: *Diabetes Mellitus, Support Vector Machine, Gestational diabetes, decision tree, Random Forest.

1. INTRODUCTION

Data mining, usually referred to as knowledge discovery using databases, is the act of locating interesting and valuable relationships and patterns within vast amounts of data. Large digital collections, or so-called data sets, require analysis using statistics and AI methods (such as neural networks and machine learning). Data mining is used in a number of sectors, including government security, finance, retail, and science research (astronomy, medical) (detection of criminals and terrorists). Data mining is more crucial than ever in the current health care context. Numerous complex reasons (data overload, early illness diagnosis and/or prevention, evidencebased care, and the reduction of hospital mistakes) may be used to promote the use of data mining in the health sector. Non-invasive diagnosis and decision-making, public health policy, and increased financial efficiency and cost savings) In data mining, the disease prediction is significant. Data mining may forecast a variety of illnesses, including hepatitis, lung cancer, liver disorders, breast cancer, thyroid conditions, diabetes, etc. The Diabetes forecasts are examined in this essay. Diabetes mellitus generally comes in four different forms. They are diabetes of the Type 1, Type 2, gestational, and congenital varieties.

Diabetes

One of the worst illnesses is diabetes. Obesity, a high glucose level, and other factors can cause diabetes. It alters the function of the hormone insulin, which causes crabs to have an irregular metabolism and raises blood sugar levels. When the body does not produce enough insulin, diabetes develops. The term “diabetes”, often known as Diabetes Mellitus, refers to a group of ailments that affects how your body converts food into energy

Type 1 Diabetes: If you have Type 1 diabetes, your pancreas either doesn't generate any insulin at all or produces very little of it, which is insufficient to allow the blood sugar to enter your cells and be utilized as energy. As a result, your blood sugar levels are abnormally high. Unusual thirst, frequent urination or the need to pee, high levels of weariness, unusual levels of hunger, and other symptoms can all be signs of type 1 diabetes. hazy vision, Loss of weight, wounds or bruises that heal more slowly.

Type2 Diabetes: In most type 2 diabetics, some insulin is typically produced by the pancreas. However, either it is insufficient or your body is not properly using it. Fat, liver, and muscle cells are typically affected by insulin resistance, which is when your cells don't react to insulin. Type 2 diabetes is frequently more manageable than type 1. The tiny blood arteries in your kidneys, nerves, and eyes are particularly vulnerable, and it still has the ability to badly impair your health. If you have type 2, your risk of heart disease and stroke rises.

Gestational Diabetes: Insulin resistance is sometimes a side effect of pregnancy. Gestational diabetes is the term used if this develops. In late or middle pregnancy, doctors frequently detect it. Gestational diabetes must be controlled in order to protect the growth and development of the baby since the mother's blood sugar levels are

transferred to the developing child through the placenta. The baby is more at danger from gestational diabetes than the mother is. A newborn may have an extraordinary prenatal weight growth, breathing difficulties at delivery, or a higher chance of developing obesity and diabetes in later life. A huge baby may require a caesarean section, or the mother may suffer injuries to her heart, kidney, nerves, and eyes.

Pre-diabetes: Technically, pre-diabetes is not a specific type of diabetes. The situation is really one where a person has raised blood sugar levels but not quite high enough to be classified as having type 2 diabetes. You have a higher risk of developing type 2 diabetes and heart disease if you have pre- diabetes. These hazards can be decreased by increasing your exercise and lowering additional weight—even only 5% to 7% of your body weight.

The majority of diabetes symptoms are similar in both sexes. Constant thirst, frequent urination, weariness, lightheadedness, and weight loss are some of these basic symptoms. Losses of muscular mass and vaginal thrush are signs that are more frequently seen in males. In addition, women frequently encounter symptoms including polycystic ovarian syndrome, vaginal yeast infections, and urinary tract infections. Diabetes can cause a wide range of significant health consequences if it is not properly treated. These include renal disease, amputation, neuropathy, retinopathy, and cardiovascular disease. Due to nerve, muscle, and blood artery damage, erectile dysfunction occurs in 45% of diabetic men. However, heart disease, renal disease, and depression are significantly more common in women. Overall, this makes it far more dangerous to women's lives than it is to men's lives. Menopause is another challenge faced by diabetic women. Diabetes and this hormonal fluctuation together may cause blood sugar levels to rise even higher, weight to gain, and sleep issues. Thus, major difficulties may develop further as a result, aggravating earlier health problems. Overall, diabetes can strike men at lower BMIs, along with extra issues including erectile dysfunction and loss of muscle mass. The decline in testosterone that occurs in males as they age is one potential explanation.

LITERATURE REVIEW

[1] Priyanka Sonar in “Diabetes Prediction Using Machine Learning Approaches” article, used Decision Tree, Support Vector Machine, Naive Bayes, and ANN. The data set used is the global Data set. The data set has seven sixty-eight instances and nine features. The demo database is the source and the target is a copy of production. The algorithm used for the classification task is Artificial Neural Network, Decision Tree, Support Vector Machine Classifier, Naive Bayes Classifier, and Machine learning Matrix. SVM is the best method to use when we don't have any idea about a dataset. Understanding the decision tree will be very easy. Naive Bayes is used to handling the missing values by ignoring estimation calculation. ANN gives good predictions and is easy to implement.

[2] In the article "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus" Md. Faisal Faruque accomplished their objective, the study methodology entails a few stages, including the gathering of a diabetic dataset with the pertinent patient variables, preprocessing the numeric value attributes, application of various machine learning classification approaches, and corresponding prediction analysis using such data. From MCC, we obtained the data set. The dataset consists of several characteristics or risk factors associated with diabetes mellitus in 200 people. The characteristics are age, sex, weight, polyuria, water intake, and blood pressure at eight. As a result, we selected four well-known machine learning algorithms for the study: Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbor (KNN)and, decision tree (DT), on adult population data to predict Diabetic Mellitus.

[3] In "Comparison of Machine Learning Algorithms in the Predicting the Onset of Diabetes," Mahmood ABED published a study. In the suggested technique, many machine learning algorithms were examined for tasks including diabetes diagnosis. Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), and Classification Naive Bayes (CNB) Classifier are among the techniques employed. They also took into account eight key factors: age, body mass index, number of pregnancies, plasma glucose concentration within two hours, Triceps of Skin Fold Thickness, diastolic blood pressure, and serum insulin within two hours. Since the MATLAB Machine Learning Toolbox has so many built-in features, including trainer, fitc- nb, fitc- knn, fitc- svm, and fitc- discr, they chose it as the platform for their investigation.

[4] Sudhansh Sharma, Bhavya Sharma, in this article “ EDAS Based Selection of Machine Learning Algorithm Based On Diabetes Detection”. Outliers are handled using the winsorization technique. Class-wise mean is the best technique for comparing the data sets than any other technique. The proposed System contains a dataset that is divided into ten parts. In n datasets, one dataset is used for testing and the remaining nine are used for training. The performance is compared with various methods based on: Accuracy, Sensitivity, and Specificity. The dataset is tested with various algorithms: Naive Bayes, Support Vector

machines, and LR. The classifiers are applied to data sets. This approach is quite successful.

[5] In the article "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", Nahla H Bharakat, Andrew P Bradley Support Vector Machine is used in this method. The linear hyperplane divides negative and positive data sets. Pre Pruning techniques are used to prune the data.SVM is tested by labeled data. The data is collected from different people of different ages, gender, different BMI, Blood pressure, Glucose level before eating, and Glucose levels after fasting. To determine if the difference in AUC between the SQRexSVM and the eclectic approaches is statistically significant, a large sample z-test was performed. a hypothesis shows that there is no difference in measured AUC between the two approaches cannot be rejected ($p > 0.05$). Furthermore, the differences in AUC between the SVM and those of the eclectic and the SQRex-SVM approaches are not statistically significant ($p > 0.05$).

[6] Prabhu, S. Selvabharathi in the manuscript "Deep belief neural network Model for Prediction of Diabetes Mellitus", has three phases. The three phases are Preprocess, Pre-training, and fine training. Before we process we need a Diabetes data set. After the data set is taken it will pass through three phases. It uses min and max normalization. DBN network is constructed. Data sets are trained using the DBM network. As the next step, they will be validated. NN-FF classification is applied to the data set. Randomly initialize all weights and biases in the network. Check for errors. Resolve those errors if there are. It gives output effectively. Here we used two values TP and FN. The false negative is FN, whereas the genuine positive is TP.

[7] "Early prediction of Diabetes Mellitus using Machine learning", GauravTripathi, Rakesh Kumar, In this classification is mainly used in different platforms like pattern recognition, it can classifies the data in to many number of classes. To build a model there is a procedure, we need to follow those procedure ,The procedures were Dataset, Data pre-processing, Algorithms used in this predictive analysis is linear discriminant analysis, k-nearest neighbor, Support vector machine ,Random forest. In this we used pima Indian Diabetes Database. There are many divergences occurred those are removed by pre-processing and it can keep the dataset clean. In this we use four algorithms to prepare a model, those four algorithms help us to predict the diabetes. The best suitable method is Random forest because it gives appropriate accuracy results.

[8] In this they have taken nearly 7 major steps for implementing their algorithm, namely Preprocess the input dataset for diabetes disease in WEKA tool, Perform percentage split of 70% to divide dataset as Training set and Test set, Select the machine learning algorithm i.e, Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm, Build the classifier model for the mentioned machine learning algorithm based on training set, Test the Classifier model for the mentioned machine learning algorithm based on test set, Perform Comparison Evaluation of the experimental performance results obtained for each classifier. Determine the highest performing algorithm after analysis based on various metrics. The proposed classifier model has been built using WEKA tool and based on successful execution of each step we can give the experimental results. For giving the accurate results they are using confusion matrix which includes Actual Class, Predicted Class, True-Positive, FalsePositive, True-Negative. At last by using four classifiers Naive Bayes, Support Vector Machine, Random Forest and Simple CART to predict the results.

2. SYSTEM REQUIREMENTS SPECIFICATION

Software Requirements

· Operating System: Windows OS

· Libraries: Keras, Tensor Flow, Numpy, ScikitLearn, Matplotlib, OpenCV

· Editor: Colab Notebook

· Technologies: Python

Hardware Requirements

· Processor: i3

· RAM: 4GB

· Hard-disk: 100GB and above

Functional Requirements

· Python3.6.2orlater, PIP, and NumPy for Windows

· Pip

· Numpy

· Pandas

· Anaconda

· Jupyter Notebook

3. METHODOLOGY

The proposed procedure is divided into various stages and each stage is explained in detail.

· Dataset as input

· Cleaning the dataset

· Splitting Data for Training and Testing

· Predict the diabetes using Machine Learning Techniques

· Compare and get final output

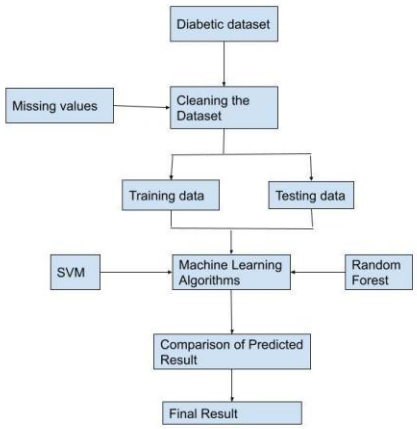

Fig1: Flowchart of proposed methodology

The notion of machine learning has swiftly gained popularity in the healthcare business. Predictions and analyses produced by the research community on medical data sets aid in illness prevention by guiding proper treatment and safeguards. Machine learning algorithms are the sorts of algorithms that can assist in decision making and prediction. We also explore several machine learning applications in the medical industry, with an emphasis on diabetes prediction using machine learning. It is likely that models trained on the same data will perform poorly in real-world scenarios because they will be over-fitted. In order to avoid this problem, it is necessary to divide the data into two parts: a training set and a test set. Normal practice is an 80/20 split.In this, we used the train test split() method from the scikit-learn package to divide the dataset into training and testing sets. 0.2 is the specified sample size.

The input dataset in diabetes contains the following fields: Pregnancy- The total number of pregnancies, Glucose- A 2-hour oral glucose tolerance test's plasma glucose levels, blood pressure or diastolic blood pressure (mm Hg), Thickness of the triceps skin fold (mm), Insulin- 2-hour serum insulin (mu U/ml), BMI- body mass index (weight in kg/(height in m)2), and other terms as applicable. diabetes pedigree function in years: (years), Class variable as a result (0 or 1).

The Dataset of diabetic patients:

	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesF	Age	Outcome
2	11	143	94	33	146	36.6	0.254	51	1
3	10	125	70	26	115	31.1	0.205	41	1
4	7	147	76	0	0	39.4	0.257	43	1
5	1	97	66	15	140	23.2	0.487	22	0
6	13	145	82	19	110	22.2	0.245	57	0
7	5	117	92	0	0	34.1	0.337	38	0
8	5	109	75	26	0	36	0.546	60	0
9	3	158	76	36	245	31.6	0.851	28	1
10	3	88	58	11	54	24.8	0.267	22	0
11	6	92	92	0	0	19.9	0.188	28	0
12	10	122	78	31	0	27.6	0.512	45	0
13	4	103	60	33	192	24	0.966	33	0
14	11	138	76	0	0	33.2	0.42	35	0
15	9	102	76	37	0	32.9	0.665	46	1
16	2	90	68	42	0	38.2	0.503	27	1
17	4	111	72	47	207	37.1	1.39	56	1
18	3	180	64	25	70	34	0.271	26	0
19	7	133	84	0	0	40.2	0.696	37	0
20	7	106	92	18	0	22.7	0.235	48	0
21	9	171	110	24	240	45.4	0.721	54	1
22	7	159	64	0	0	27.4	0.294	40	0
23	0	180	66	39	0	42	1.893	25	1
24	1	146	56	0	0	29.7	0.564	29	0
25	2	71	70	27	0	28	0.586	22	0

diabetes

Fig2: Diabetes dataset

The result is achieved in the following steps:

- Read the diabetes data set
- The structural data of the data set were examined using exploratory data analysis. The dataset's variable kinds were investigated. The dataset's size information was obtained. The data set's 0 values represent missing values. These 0 values were primarily changed to NaN values. The data set's descriptive statistics were looked at.
- Section 3 of data preprocessing; df for The median values of whether each variable was unwell or not were used to NaN values for missing data should be filled in. The LOF identified the outliers and eliminated them. The robust technique was used to normalise the X variables.
- Cross Validation Score was determined during Model Building utilising machine learning models such as Logistic Regression, KNN, SVM, CART, Random Forests, XGBoost, and LightGBM. Later hyperparameter modifications to Random Forests, XGBoost, and LightGBM designed to boost Cross Validation value
- The prediction of diabetes is resulted accurately by random forest algorithm. The final output is attained by comparing the all methodologies.

Algorithms used:

a) **Support Vector Machine**

A method of supervised machine learning called a Support Vector Machine (SVM) can be used to solve classification and regression problems. The majority of its applications come from classification problems. As part of this algorithm, each data item is plotted as a point in n-dimensional space (where n represents how many features you have) with each feature's value being the coordinate value. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the figure below). Kernels are used in practice to implement the SVM algorithm. In linear SVM, hyper-planes are learned using linear algebra, which is beyond the scope of this introduction. A powerful insight is that linear SVM can be rephrased by taking the inner product of any two given observations, rather than the observations themselves. The inner product between two vectors is the sum of the multiplication of each pair of input values. The equation for predicting a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

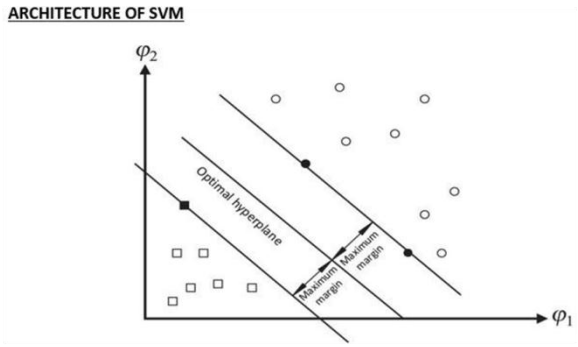

Fig3: Support Vector Machine

b) **Random Forest**

Now that computing power has increased, we may select algorithms that carry out extremely complex calculations. The algorithm "Random Forest" is one example. With a decision tree (CART) model, random forest is comparable to a bootstrapping process. Let's say that there are 1000 observations in the entire population across 10 variables. With various samples and various beginning factors, Random Forest attempts to construct several CART models. A CART model, for instance, will require a random sample of 100 observations and 5 randomly selected beginning variables. After doing the procedure, let's say, ten times, it will generate a final forecast based on each observation. Each forecast influences the final one. The mean of all previous predictions can be used as the final forecast. Some decision trees may predict the proper output, while others may not, since the random forest mixes numerous trees to forecast the class of the dataset. But when all the trees are combined, they forecast the right result. In comparison to other algorithms, it requires less training time. Even with the enormous dataset, it operates effectively and predicts the outcome with a high degree of accuracy. Accuracy can be kept even when a sizable portion of the data is missing.

4. RESULT

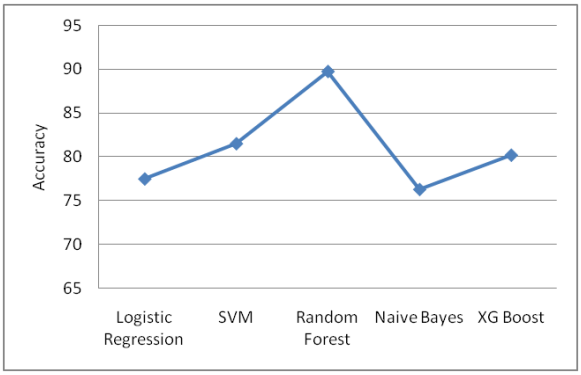

Fig3: Graphical representation of outcomes

The random forest method yields the most accurate result. The graph displays the general effectiveness of every strategy on the same set of data. Due to the lack of data, Logistic Regression and Naive Bayes have relatively low prediction rates. When the data is missing, the prediction is inaccurate. Random forest achieves the ultimate high prediction rate because it retains good accuracy even when a significant amount of the data is missing.

Methodology	AUC	ROC
Logistic Regression	0.80	0.77
Support Vector Machine	0.79	0.71
KNeighbors	0.66	0.58
Random Forest	0.81	0.75
Naive Bayes	0.74	0.69

The proposed approach built in python and make use of various classification and ensemble algorithms. These technique are common Machine Learning techniques usedc to get the maximum accuracy out of data. We can see from the competition. Overall, make predictions and achieve high performance accuracy, we applied the best machine learning approaches. The outcome of these machine learning techniques is shown in Figure.

5. CONCLUSION

Machine learning algorithms and data mining algorithms in the medical field can identify hidden patterns in medical data. The systems can be used to analyze critical clinical parameters, predict diseases, forecast medical tasks, extract medical knowledge, support therapy planning and maintain patient records. There have been several algorithms proposed for predicting and diagnosing diabetes. These algorithms provide more accuracy than the available traditional systems. We tried and optimized every algorithm and we found the RANDOM FOREST algorithm to be the most suitable for over applications. Future research may involve the prediction or diagnosis of other diseases that use the developed system and the machine learning classification methods. Other machine learning algorithms can be added to the work to improve and expand it for the automation of diabetes analysis.

6. REFERENCES

[1] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

[2] M. F. Faruque, Asaduzzaman and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-4, doi: 10.1109/ECACE.2019.8679365.

[3] M. Abed and T. İbrıkçı, "Comparison between Machine Learning Algorithms in the Predicting the Onset of Diabetes," 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 2019, pp. 1-5, doi: 10.1109/ IDAP.2019.8875965.

[4] S. Sharma , "EDAS Based Selection of Machine Learning Algorithm for Diabetes Detection," 9th International Conference System Modeling and Advancement in Research Trends (SMART) 2020, Moradabad, India, 2020, pp. 240-244, doi: 10.1109/ SMART50582.2020.9337118.

[5] N. Barakat, A. P. Bradley and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pp. 1114-1120, July 2010, doi: 10.1109/TITB.2009. 2039485.

[6] P. Prabhu and S. Selvabharathi, "Deep Belief Neural Network Model for Prediction of Diabetes Mellitus," 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC), Singapore, 2019, pp. 138-142, doi: 10.1109/ICISPC.2019.8935838. [7] G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 1009-1014, doi: 10.1109/ICRITO48877.2020.9197832.

[8] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp.1-6, doi: 10.1109/ICCUBEA.2018 .8697439.

[9] S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 141-146, doi: 10.1109/ICACCS51430.2021.

[10] C. Huang, G. Jiang, Z. Chen and S. Chen, "The research on evaluation of diabetes metabolic function based on Support Vector Machine," *2010 3rd International Conference on Biomedical Engineering and* Informatics, Yantai, China, 2010, pp. 634-638, doi: 10.1109/BMEI.2010.5640041.

[11] R. Deo and S. Panigrahi, "Performance Assessment of Machine Learning Based Models for Diabetes Prediction," 2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT), Bethesda, MD, USA, 2019, pp. 147-150, doi: 10.1109/HI-POCT45284.2019.8962811.

[12] M. Ayad, H. Kanaan and M. Ayache, "Diabetes Disease Prediction Using Artificial Intelligence," 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, 2020, pp. 1-6, doi: 10.1109/ACIT50332.2020.9300066.