International Research Journal of
SCIENCE, TECHNOLOGY, EDUCATION, AND MANAGEMENT
P-ISSN: 2799-063X | E-ISSN: 2799-0648
ISO 9001:2015 Certified

Volume 2, No. 2 | July 2022

Support vector machine classification learning algorithm for diabetes prediction

Renas Rajab Asaad
Department of Computer Science, College of Science, Nawroz University, Kurdistan-Region, Iraq
Renas.Rekany@Nawroz.edu.krd

ABSTRACT

Diabetes is actually one of the primary causes of human mortality. Diabetes is an intense disease affecting various parts of the human body. Diabetes can rise long-range complications including, renal failure and cardiac failure. It is therefore imperative that diabetes be diagnosed in a timely manner to people all over the world. This study develops a method for diabetic classification using machine learning techniques. In this study, Support Vector Machine (SVM) is employed to classify the diabetic disease into two classes based on its different functions, namely, linear, polynomial, and sigmoid functions. The evaluation performance of this study is performed before and after applying the pre-processing stage using different standard criteria. The higher results were obtained by polynomial function 83.77% for accuracy, 86.07% for sensitivity, and 81.97% for specificity. Finally, a comparison between this study and some of the previous studies was addressed, based on the comparison it is shown that this study has a better ability to classify diabetic disease than previous studies.

ARTICLE INFO

Received : April 2, 2022
Revised : May 17, 2022
Accepted : June 14, 2022

KEYWORDS
Classification, Data analysis, Diabetic, Feature extraction, Machine learning, PIMA dataset, Pre-processing, SVM technique,

Suggested Citation (APA Style 7th Edition):

Asaad, R.R. (2022). Support vector machine classification learning algorithm for diabetes prediction. *International Research Journal of Science, Technology, Education, and Management*, 2(2), 26-34. https://doi.org/10.5281/zenodo.6975670

INTRODUCTION

Even before technology, health has always been a priority. Because healthcare has grown so much, there is sufficient space for research in this field. It is critical to upgrade existing healthcare technology by adopting medical information digitalization, both in terms of patient-provided data and medical findings supplied by advanced equipment. We are faced with the difficult problem of deciphering and understanding the massive data gathered as a result of this digital revolution. As a result of a large amount of data, big data analytics comes to the rescue (Mir & Dhage, 2018).

For assisting medical specialists, numerous decision support systems based on various data mining methods are presented. The accuracy of the decision support system is a measure of its effectiveness. As a result, the primary goal of developing a decision support system is to accurately predict and diagnose a specific disease (Vijayan, & Anjali, 2015). Diabetes is a chronic disease that is also referred to as the "silent disease." Diabetes is defined by the World Health Organization (WHO) as a disease in which the body is unable to efficiently utilize the energy provided by the food it consumes. Furthermore, the condition develops when the pancreas produces the hormone insulin, which aids the body's consumption of sugar and fat, as well as the storage of part of it (Edeh et al., 2022). Two factors can cause diabetes: (1) The pancreas produces insufficient insulin, or the body produces inadequate insulin. This type of diabetes affects only about 5% to 10% of people with diabetes (Type-1). (2) The insulin produced has no effect on the cells (Type-2) (Kumari & Chitra, 2013).

The day requires to detection and diagnosis of diabetes at an early stage. Diagnosis of diabetes and interpretation of diabetes data is a significant categorization issue. Machine learning is an artificial intelligence subfield of computer science that is used to create intelligent devices that can learn without being programmed. Pattern recognition and computational learning theory have been the foundations of machine learning. There are three types of classification algorithms used in machine learning: supervised, unsupervised, and reinforcement. Machine learning algorithms such as Naive Bayes, logistic regression, random forest, and the KNN algorithm were utilized to achieve this. Diabetes is characterized in three ways, and the performance is compared based on accuracy. (Nibareke, & Laassiri, 2020) (Swarna et al., 2020). SVM has been used to recognize and classify patterns in a variety of applications, including handwriting identification, face detection, data analysis, and more. The purpose of SVM models is to find the hyperplane with the most margins between the specified target classes and classify them as quickly as possible. The major goal of this project is to create a model for diabetic categorization based on the use of a Support Vector Machine (SVM) with various features.

LITERATURE REVIEW

Diabetes is predicted and diagnosed using a variety of classification and clustering algorithms. Various machine-learning techniques, such as Naive Bayes, KNN algorithm, random forest, and logistic regression are used to predict diabetes based on a data set (Swarna et al., 2020). The study's main goal is to use big data techniques and a machine learning model to track diabetes disease progression. The authors can do this by using some matrices to select a more realistic model. This research evaluates the effectiveness of four machine learning models in terms of predicting diabetic illness. The Naive Bayes classifier has a greater precision of 78 percent, according to the trial results (Katarya & Jain, 2020) Using the Python data manipulation tool, I used the machine learning technique to find patterns and risk factors in the Pima Indian diabetes dataset. To determine whether a patient is diabetic or non-diabetic. Based on the experimental data, this study used various machine learning classifiers for diabetic prediction, and the greater accuracy reached was 84 percent. The emergence of metabolic syndrome and diabetes was a concern for Perveen et al. (Smys, 2019). The training set was balanced using k-methods sampling, and the Naive Bayes and J48 (C4.5) decision tree models were employed for prediction. The goal of Sneha et al. (Sneha & Gangil, 2019) was to determine the best classifier. The dataset comes from UCI's machine learning library. There are 2500 data pieces and 15 attributes in all. 768 data items are chosen for testing using categorization algorithms. SVM, DT, Naive Bayes, KNN, and Random Forest algorithms are evaluated using the quick miner tool. Naive Bayes beat other classifiers in the suggested technique, reaching an accuracy of 82.30 percent. Yichuan Wang et al.

(Wang et al.,2018) suggested a data analytics structure for the healthcare industry, using which they defined five big data analytics entities: pattern analysis, unstructured data analysis, decision support, predictive analytics, and traceability.

The proposed model in (Giveki et al., 2012) is divided into three steps: first, the PCA is used to define an ideal subset of features from a collection of all features. Second, by weighting distinct characteristics based on their value, mutual information is employed to create (Feature Weighted Support Vector Machines). Finally, because parameter selection is very important for SVM rating accuracy, MCS is used to find the best parameter values. They proposed examining the DT method and rating it based on accuracy in this paper (Alghurair, 2020). For low-volume sample answers, decision trees tend to match rapidly and have poor forecast accuracy.

Various supervised learning methods were utilized in the research such as CT, SVM, k-NN, NB, RF, NN, AB, and LR, and generated the training dataset and testing dataset using k-fold cross-validation with k = 10 (Edeh et al., 2022). The accuracy resulted in a score of 76.80%. Using the Ridge-Adaline Stochastic Gradient Descent Classifier (RASGD), the study in (Deepa et al., 2021) intends to develop an artificial intelligence-based intelligent system for earlier disease prediction. The proposed scheme's outcomes were compared to current machine learning algorithms like support vector machines and logistic regression approaches. The RASGD intelligent system achieves a 92 percent accuracy rate, which is better than the other classifiers tested (Nibareke & Laassiri, 2020).

Proposed method

A brief explanation of the progress of the technology used is provided in this section. The suggested classifier model primarily warns diabetic patients and incorporates diabetes data into the data set. We begin by gathering the contextualized Pima Indian diabetes dataset. Exploratory data analysis is used to better understand the sources from which our data is gathered. The next step is to preprocess our data, which entails cleaning our dataset by removing any duplicate, missing, or odd entries. The next step is to choose the models that will be used to train our data and fit the model. The models will then be compared using various performance criteria such as accuracy, sensitivity, and specificity. Figure 1 depicts the proposed methodology and the processes involved in its implementation.

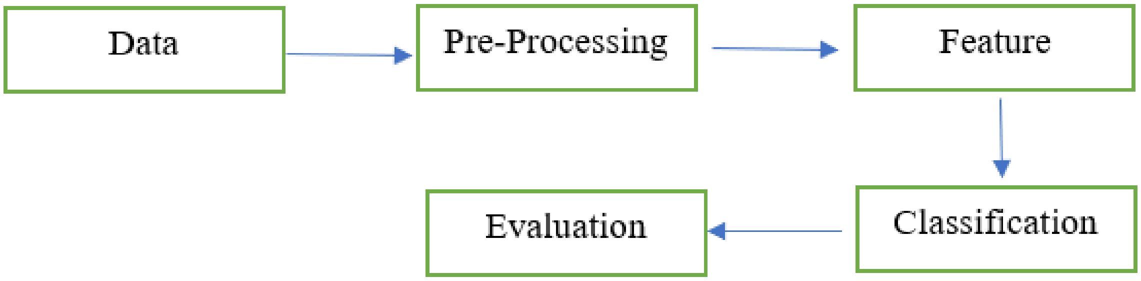

Figure 1. Proposed methodology

Pre-processing

By dividing a score's variance by the standard deviation of a data set, a Z-score or standard score is used to standardize scores on a constant scale. It determines how far a given data is from the mean in terms of standard deviations. The value is between (-1, 1), with a negative Z-score for values less than the mean, a positive Z-score for values greater than the mean, and a zero Z-score for the mean. The mean () and standard deviation () of the attributes can be used to determine the Z-score normalization (Padmavathi & Sumathi, 2019).

Z-Score $= \frac{x-u}{\sigma}$ (1)

Machine learning

Any sort of data can be used in a supervised learning strategy. Classification learns from the input data and then classifies new data based on it. This method aids in the identification of class labels into which fresh data can be inserted. Different labels and groups can be produced using the clustering method based on similarities in the data set. Classification is a data mining process that divides data into categories to aid in more accurate forecasting and analysis. It's one of the data mining techniques for analyzing massive datasets. (Tan et al., 2016). It's utilized to create patterns that accurately characterize the data set's essential data classes. The target classes for each of the current data instances are predicted using classification techniques. The goal of classification algorithms is to find correlations between attributes that can be used to predict the outcome. They examine the data and make a prediction. Data mining categorization tasks are commonly employed in the healthcare industry (Selvaraj et al., 2020).

Support vector machine

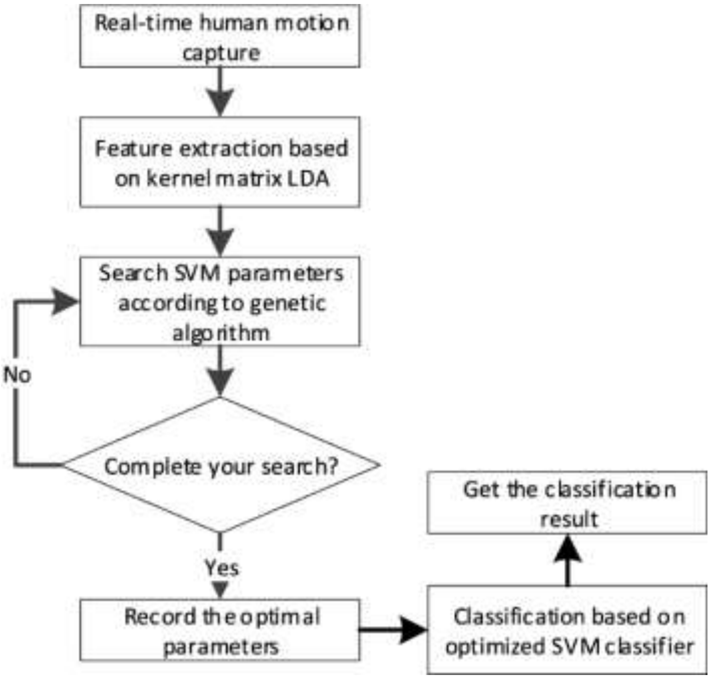
Real-time human motion capture
Feature extraction based on kernel matrix LDA
Search SVM parameters according to genetic algorithm
No
Complete your search?
Yes
Record the optimal parameters
Classification based on optimized SVM classifier
Get the classification result

Figure 2. SVM technique

SVM technique

It is critical to train and test a model in order to achieve a decent prediction result. The training dataset is a collection of learning sets that must be used to train learning and prediction algorithms (Vishwanathan & Murty, 2002). The test set is only used to evaluate the performance of the classifiers that have been chosen. It's solely there to help with classifier testing. If a model outperforms both datasets, the expected accuracy is higher (Yue et al., 2003).

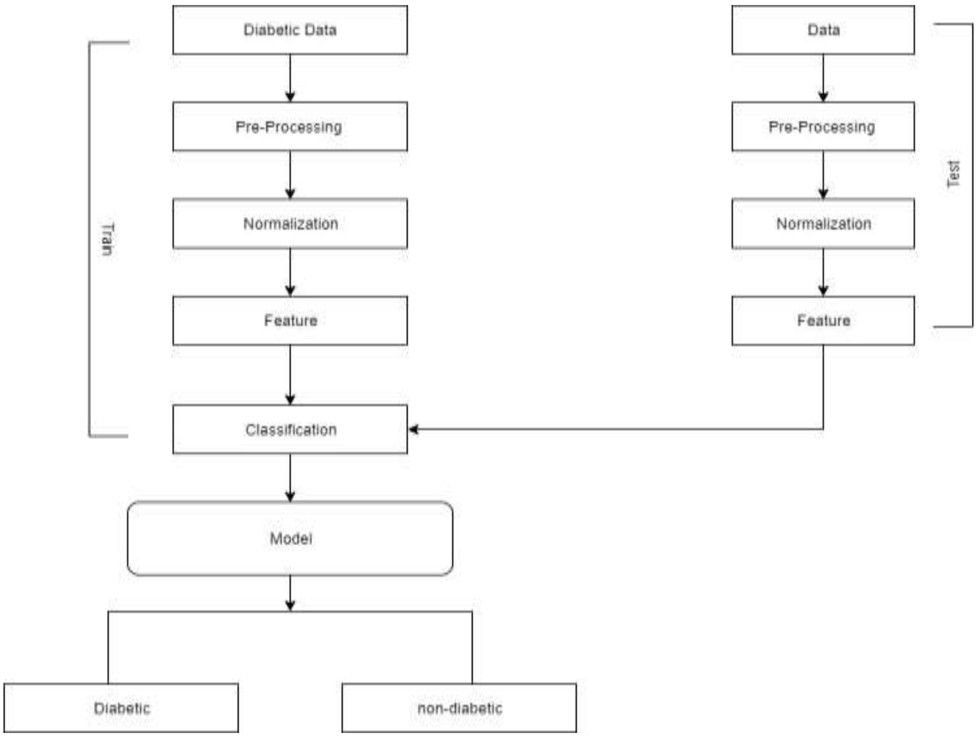

Figure 3. Developed model

Experimental result

Dataset

The diabetes csv dataset is obtained from the Pima India diabetes database. It contains eight properties that act as diabetes markers. Based on 768 cases, they test for the presence of diabetes in the patient. In patients, this standard might take several forms. The physical signs of the data set are shown in the table below. The 8 attributes of the diabetes dataset are quickly described in Table I. (Swarna et al., 2020). For this research study, the Panda-assisted diabetes database file was read, which comprises 768 records and 8 medical predictor features for input and 1 target variable output, with 1 for 'yes' and 0 for 'no' diabetes. Figure 4 shows that out of 768 Pima Indian women, 65.1 percent had not been diagnosed with diabetes. While 34.90 Pima Indian women out of 768 have diabetes (Zhu & Blumberg, 2002) (Senthilkumar et al., 2018).

Table 1. Attributes of dataset

No.	Attributes	Less Than
1	Pregnancies	17
2	Glucose	199
3	Blood Pressure	122
4	Skin thickness	99
5	Insulin	846
6	BMI	67.1
7	Diabetes Pedigree Function	2.42
8	Age	81

Confusion matrix

A confusion matrix is used in machine learning to evaluate the performance of a classification algorithm. The confusion matrix is a table with rows representing the actual class and columns representing the predicted class.

Table 2. Confusion matrix structure

Confusion Matrix Structure			
Total No. of Instances		Predicted Class	
		No a=tested negative	Yes b=tested positive
Actual Class	No a=tested negative	True Negative	False Negative
	Yes b=tested positive	False Negative	True Negative

Certain terminology that appears in the general Confusion Matrix Structure is described below. This terminology will be further used for the Performance Evaluation of each classifier.

- True Positive: No. of instances predicted positive and are actually positive
- False Positive: No. of instances predicted negative and are actually negative
- True Negative: No. of instances predicted positive but are actually negative
- False Negative: No. of instances predicted negative but are actually positive

Accuracy, sensitivity, specificity, and error rate are among the performance measures that are assessed. The proportion of correct forecasts to the total number of forecasts is referred to as accuracy. Equation contains the formula (1). The fraction of positive samples that are tested positive is known as sensitivity. It's also known as a true positive rate. Equation contains the formula (2). The fraction of negative samples that are tested negative is known as specificity. It's also known as a true negative rate. Equation contains the formula (3).

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$\frac{TP}{TP+FN} \tag{3}$$

$$\frac{TN}{TN+FP} \tag{4}$$

This section presents the performance evaluation results of this study based on different matrices' accuracy, sensitivity, and specificity based on the SVM classifier. The evaluation has been performed in two-stage before applying to pre-processing and after applying to pre-process. The following tables (2 and 3) represent the obtained results for all classifiers after building on the PIAM diabetic dataset. According to the classification results obtained in Table 2 before applying the pre-processing, sigmoid has obtained lower classification results by 65.78% of accuracy as well as in sensitivity and specificity. The obtained accuracy for the linear kernel was better than the sigmoid by obtaining the 72.45% as well as the sensitivity and specificity were better. However, the polynomial was obtained best accuracy, sensitivity, and specificity by 80.03%, 82.22%, and 79.99%.

Table 4. Obtained results for diabetic classification before pre-processing

	Accuracy	Sensitivity	Specificity
Linear	72.45	75.21	70.01
Ploy	80.03	82.22	79.99
Sigmoid	65.78	69.27	64.81

On the other hand, after applying the pre-processing stage and then building a classifier to classify PIMA diabetic dataset, the evaluation results were improved. According to Table 3 as it shows the obtained results after pre-processing. The results of all classifiers were improved, and the accuracy, sensitivity, and specificity obtained by sigmoid were 68.04%, 72.39%, and 67.01. Also, the results obtained by the linear were better than sigmoid by obtaining 75.29% accuracy, 79.11% sensitivity, and 73.61% specificity. Finally, the best classification results for this study were obtained by polynomial where the accuracy was 83.77%, sensitivity 86.07%, and specificity 81.97%.

	Accuracy	Sensitivity	Specificity
Linear	75.29	79.11	73.61
Ploy	83.77	86.07	81.97
Sigmoid	68.04	72.39	67.01

Table 5. Obtained results for diabetic classification after pre-processing

SVM results

Support Vector Machine (SVM) with it is different kernels have been built as classifiers for classifying PIMA diabetic dataset in this study. As has been shown before the presented algorithm has been tested and evaluated based on using widely used matrices in this research field (accuracy, sensitivity, and specificity). The testing results obtained by this study were in two stages before and after pre-processing. The higher testing accuracy of the diabetic dataset was 80.03% and the testing accuracy obtained after pre-processing was 83.77%. From the experimental results, it is noticed that the accuracy performance has been improved after applying the pre-processing. Moreover, this section compares obtained performance evaluation accuracy with some of the previous studies based on using the same dataset. Table 4 shows the comparisons including this study and some of the previous studies. Table 4 shows the presented comparison between this study and previous studies.

Reference	Method	Accuracy
(Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. ,2018)	Levenberg-Marquardt	82%
(Choubey, D. K., & Paul, S. ,2017)	GA_RBF NN	77.4%
(Ateeq, K., & Ganapathy, G. ,2017)	MPSO-NN	81.8%
This study	Poly- SVM	83.77%

Table 6: Comparison between this study and previous studies

CONCLUSION

One of the world's most important modern medical challenges is early diabetes detection. Trials were performed on the Pima Indian patient database. Prediction analysis is how the future user predicts the basis for current situations. During this study, a survey was conducted on the results obtained from data testing using SVM technology with a different kernel function. The result of this was poor compared to other previous algorithms, where it obtained the highest accuracy by up to 83%.

REFERENCES

Ateeq, K. & Ganapathy, G. (2017). The novel hybrid Modified Particle Swarm Optimization–Neural Network (MPSO-NN) Algorithm for classifying Diabetes. *International Journal of Computational Intelligence Research*, 13(4), 595-614.

Bhoia SK, Pandab SK, Jenaa KK, Abhisekhc PA, Sahood KS, Samae NU, et al. 'Prediction of Diabetes in Females of PimaIndian Heritage: A Complete Supervised Learning Approach'. *Turk J Comput Math Educ*. (2021) 12:3074–84.

Choubey, D.K., & Paul, S. (2017). GA_RBF NN: a classification system for diabetes. *International Journal of Biomedical Engineering and Technology*, 23(1), 71-93.

Deepa N, Prabadevi B, Maddikunta PK, Gadekallu TR, Baker T, Khan MA, et al. An AI-based intelligent system for healthcare analysis using Ridge- Adaline Stochastic Gradient Descent Classifier. *J Supercomput*. (2021) 77:4– 16. doi: 10.1007/s11227-020-03347-2

Edeh, M.O., Khalaf, O.I., Tavera, C.A., Tayeb, S., Ghouali, S., Abdulsahib, G.M., & Louni, A. (2022). A classification algorithm-based hybrid diabetes prediction model. *Frontiers in Public Health, 10*.

Giveki, D., Salimi, H., Bahmanyar, G., & Khademian,Y. (2012).Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search .ArXiv, 2319-7595.

Katarya, R. & Jain, S. (2020). Comparison of Different Machine Learning Models for diabetes detection. In *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)* (pp.1-5). IEEE.

Kumari, V.A. & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2), 1797-1801.

Mir, A. & Dhage, S.N. (2018). Diabetes disease prediction using machine learning on big data of healthcare. In *2018 fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-6). IEEE.

mys, S. (2019). "Survey on accuracy of predictive bigdata analytics in healthcare." *Journal of Information Technology* 1, no. 02 (2019): 77-86.

Nibareke, T. & Laassiri, J. (2020). Using Big Data-machine learning models for diabetes prediction and flight delays analytics. *Journal of Big Data*, 7(1), 1-18.

Padmavathi, M.S. & Sumathi, C.P. (2019). A New Method of Data Preparation for Classifying Diabetes Dataset. *Indian Journal of Science and Technology*, *12*(22), 1-9.

Pang-Ning Tan; Michael Steinbach; Anuj Karpatne; Vipin Kuma (2019). "Introduction to Data Mining 2nd ed", Publisher: Pearson.

Selvaraj, B., Pavithra, S.V., Rak, A.N., & Jeyaselvi, M. (2020). Prediction And Detection Of Diabetes Using Machine Learning. *International Research Journal of Engineering and Technology* (IRJET).

Senthilkumar, S.A., Rai, B.K., Meshram, A.A., Gunasekaran, A., & Chandrakumarmangalam, S. (2018). Big data in healthcare management: a review of the literature. *American Journal of Theoretical and Applied Business*, 4(2), 57-69.

Sneha, N. & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, 6(1), 1-19.

Swarna, S.R., Boyapati, S., Dixit, P., & Agrawal, R. (2020). Diabetes Prediction by using Big Data Tool and Machine Learning Approaches. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 750-755). IEEE.

Thoma, J., Joseph, A., Johnson, I., Thomas, J. (2019). Machine Learning Approach For Diabetes Prediction. *International Journal of Information Systems and Computer Sciences*, 2319-7595.

VeenaVijayan, V. & Anjali, C. (2015). Prediction and diagnosis of diabetes mellitus—a machine learning approach. Recent Adv. 2015.

Vishwanathan, S.V.M. & Murty, M.N. (2002). SSVM: a simple SVM algorithm. In Proceedings of the 2002 International Joint Conference on Neural Networks. (Vol. 3, pp. 2393-2398). IEEE.

Wang, Y., Kung, L., & Byrd, T.A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126, 3-13.

Yue, S., Li, P., & Hao, P. (2003). SVM classification: Its contents and challenges. Applied Mathematics-A Journal of Chinese Universities, 18(3), 332-342.

Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. (2018). A Feed-Forward Neural Network Model For The Accurate Prediction Of Diabetes Mellitus. *International Journal of Scientific & Technology Research*, 2277-8616.

Zhu, G. & Blumberg, D.G. (2002). Classification using ASTER data and SVM algorithms;: The case study of Beer Sheva, Israel. *Remote sensing of Environment*, 80(2), 233-240.