# **Literature review**

Diabetes or diabetes mellitus is a metabolic disorder (metabolic) in the body. This disease destroys the ability to produce insulin in the patient's body or the body develops resistance to insulin the and consequently the produced insulin cannot achieve its normal job. The main role of the produced insulin is to decrees blood sugar by different instruments. There are two key types of diabetes. In Type I diabetes, obliteration of beta pancreatic cells damage insulin construction and in type II, there is a progressive insulin confrontation in the body and ultimately may yield to the obliteration of pancreatic beta cells and faults in insulin production. In type II diabetes, it is known that genetic issues, obesity and lack of physical activity have a vital part in a person [1].

Even though the precise cause of type I diabetes is unidentified, issues that may indicate a greater risk comprise the followings [2]:

Family history**.** A person risk upsurges if his parent or sibling has history of type I diabetes.

 Environmental factors. Situations for example contact with a viral illness probably play some role in type I diabetes.

 The existence of harmful immune system cells. Occasionally family members of a person with type I diabetes are examined for the existence of diabetes autoantibodies. If a person has these autoantibodies, he/she has a chance of increased risk for evolving type I diabetes. Nonetheless not every person who has these autoantibodies gets diabetes.

 Geography. Some countries, like Sweden, have bigger rates of type I diabetes.

Environmental factors. Situations for example contact with a viral illness probably play some role in type I diabetes.

Researchers don't completely comprehend why certain people develop pre-diabetes and type II diabetes and others don't. It's sure that some factors upsurge the risk like [2]:

 Weight. The more fatty tissue you have, the more resilient a person cells to insulin.

 Inactivity. The less energetic a person is, the more a person has risk. Physical activity assists a person control of his/her weight, consumes glucose as energy and makes a person cells more sensitive to insulin.

 Family history. A person risk upsurges if his parent or sibling has history of type II diabetes.

 Race. Even though it's uncertain why, people of specific races are at higher risk.

 Age. A person risk upsurges as he/she gets older. This may be because a person has a habit to exercise less, lose muscle mass and add weight as he/she gets older. Nonetheless type II diabetes is likewise growing among children, youths and adults.

 Gestational diabetes. If a person developed gestational diabetes when she was pregnant, her risk of emerging pre-diabetes and type II diabetes far ahead upsurges. If she gives birth to a baby weighing more than 4 kilograms, she is also at risk of type II diabetes.

 Polycystic ovary syndrome. For females, having polycystic ovary syndrome increases the risk of getting diabetes.

 High blood pressure. Having blood pressure more than 140/90 millimeters of mercury (mm Hg) is connected to an augmented risk of type II diabetes.

 Abnormal cholesterol and triglyceride levels. If a person has low levels of high-density lipoprotein, or good cholesterol, his/her risk of type II diabetes is going to be higher. Triglycerides are additional type of fat passed in the blood. A person with greater levels of triglycerides has an augmented risk of type II diabetes.

A number of disease prediction models are used in medical diagnosis system which are using data mining and machine leaning techniques like Bayesian classification, Decision Tree, Regression model, Neural Network etc. We choose neural-network, logistic regression and decision tree for solving this problem.

## **Decision tree:**

Decision tree is a basic classification and regression method. Decision tree model has a tree structure, which can describe the process of classification instances based on features [3]. It can be considered as a set of if-then rules, which also can be thought of as conditional probability distributions defined in feature space and class space.

Decision tree uses tree structure and the tree begins with a single node representing the training samples [4]. If the samples are all in the same class, the node becomes the leaf and the class marks it. Otherwise, the algorithm chooses the discriminatory attribute as the current node of the decision tree. According to the value of the current decision node attribute, the training samples are divided into serval subsets, each of which forms a branch, and there are serval values that form serval branches [5]. For each subset or branch obtained in the previous step, the previous steps are repeated, recursively forming a decision tree on each of the partitioned samples [5,6].

The typical algorithms of decision tree are ID3, C4.5, CART and so on. In this study, we used the J48 decision tree in WEKA. J48 another name is C4.8, which is an upgrade of C4.5. J48[7] is a top-down, recursive divide and conquer strategy. This method selects an attribute to be root node, generates a branch for each possible attribute value, divides the instance into multiple subsets, and each subset corresponds to a branch of the root node, and then repeats the process recursively on each branch [8]. When all instances have the same classification, the algorithm stop. In J48, the nodes are decided by information gain. According to the following formulas, in each iteration, J48 calculates the information gain of each attribute, and selects the attribute with the largest value of information gain as the node of this iteration [9].

Attribute A information gain:

Gain(A)=Info(D)−InfoA(D)

Pre-segmentation information entropy:

Info(D)=Entropy(D)=−Σjp(j|D)logp(j∣∣∣d)

Distributed information entropy:

InfoA(D)=Σi=1vninInfo(Di)

## **Neural Network:**

Neural network is a math model, which imitates the animal’s neural network behaviors. This model depends on the complexity of the system to achieve the purpose of processing information by adjusting the relationship between the internal nodes (Mukai et al., 2012). According to the connections’ style, the neural network model can be divided into forward network and feedback network. In this paper, we used the Neural Pattern Recognition app in MATLAB, which is a two-layer-feed-back network with sigmoid hidden and SoftMax output neurons. The neural network structural is shown in (Figure 1).

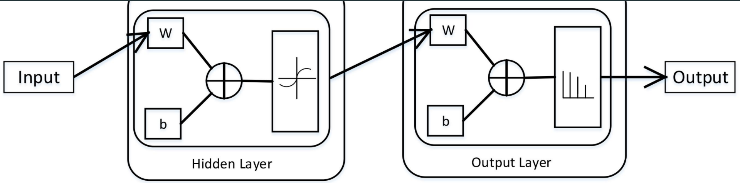


FIGURE 1. The structural of two–layer-feed-back network in MATLAB. This figure is from MATLAB, which can describe this network working principle preferably. Where, W is representation the weight and b is the bias variable.

In neural network, there are some important parts, namely input layer, hidden layer and output layer. The input layer is responsible for accepting input data. We can get the results from the output layer. The layer between the input layer and the output layer is called hidden layer. Because they are invisible to the outside. There is no connection between neurons on the same layer. In this network, the number of hidden layers set to 10, which can get a better performance. We suppose the input vector is x→, the weight vector is w→, and the activation function is a sigmoid function, then the output is:

y=sigmoid(w^→T⋅x^→)

and the sigmoid is:

sigmoid(x)=1/1+e^−x

**3.** 3 ***Logistic Regression:***

Logistic Regression is Classification algorithm commonly used in Machine Learning. It allows categorizing data into discrete classes by learning the relationship from a given set of labeled data. It learns a linear relationship from the given dataset and then introduces a non-linearity in the form of the Sigmoid function.

In case of Logistic regression, the hypothesis is the Sigmoid of a straight line, i.e,

h(x) = \sigma(wx + b) where \sigma(z) = \frac{1}{1 + e^{-z}}

Where the vector w represents the Weights and the scalar b represents the Bias of the model.

Note that the range of the Sigmoid function is (0, 1) which means that the resultant values are in between 0 and 1. This property of Sigmoid function makes it a really good choice of Activation Function for Binary Classification. Also for z = 0, Sigmoid(z) = 0.5 which is the midpoint of the range of Sigmoid function.

Just like Linear Regression, we need to find the optimal values of w and b for which the cost function J is minimum. In this case, we will be using the Sigmoid Cross Entropy cost function which is given by

J(w, b) = -\frac{1}{m} \sum\_{i=1}^{m}(y\_i \* log(h(x\_i)) + (1 - y\_i) \* log(1 - h(x\_i)))

This cost function will then be optimized using Gradient Descent.

REFERENCES

[5] World Health Organization (WHO), "Definition, Diagnosis, and classification of diabetes mellitus and its complications", part 1. WHO/NCD/NCS/2016.2, (2016).

[6] H. Temurtas, N. Yumusak and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", Expert System, vol. 36, (2009), pp. 8610–15.

[7]Quinlan, J. R. (1996a). “Bagging, boosting, and C4.5,” in Proceedings of the Thirteenth National Conference on Artificial Intelligence (Menlo Park, CA: AAAI Press), 725–730.

[8]Liao, Z. J., Wan, S., He, Y., and Zou, Q. (2018). Classification of small GTPases with hybrid protein features and advanced machine learning techniques. Curr. Bioinform. 13, 492–500. doi: 10.2174/1574893612666171121162552

[9]Quinlan, J. R. (1996a). “Bagging, boosting, and C4.5,” in Proceedings of the Thirteenth National Conference on Artificial Intelligence (Menlo Park, CA: AAAI Press), 725–730.

[10]Habibi, S., Ahmadi, M., and Alizadeh, S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob. J. Health Sci. 7, 304–310. doi: 10.5539/gjhs.v7n5p304

[11]Salzberg, S. L. (1994). C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, Inc., 1993. Mach. Learn. 16, 235–240.

[12]Kohavi, R. (1995). “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal

[13]Sharma, S., Agrawal, J., and Sharma, S. (2014). classification through machine learning technique: C4. 5 algorithm based on various entropies. Int. J. Comput. Appl. 82, 28–32