CHAPTER 2
LITERATURE REVIEW

2.1. Related Works

This section we are going to provide some of related works that have been accomplished in the past by previous researcher which has a similarity and relevancy with this research, this related works will be useful as a reference to drive this research to go through into correct process which enhancing the previous works.

Arslan, Polat, & Güneş have done their research in classifying PID using PCA and ANFIS algorithm in 2008, they have claimed 89% accuracy. In the same year, Kahramanli and Allahverdi are adopting ANN and FNN, it produces 84% as the accuracy. In 2011, Luukka used fuzzy entropy-based feature selection and similarity classifier for PID Dataset, it generates 75.97% as accuracy.

In 2013, (Beloufa & Chikh, 2013) convey that Diabetes Mellitus is one of the most dangerous debilitating disease that need to be fight worldwide, by following the rapid growth in information technology both setup their plan to fight DM by leveraging Artificial Intelligence with their proposed method called novel Artificial Bee Colony (ABC) to classify the potential of DM in someone. In summarize, they convey that ABC can be an efficient and reliable method to classify diabetes, in this research they leverage Pima Indian Diabetes (PID) datasets and it had been concluded that ABC algorithm is a powerful tool for diagnosing Diabetes Mellitus.

In the same year with (Beloufa & Chikh, 2013) launched their research, (Wang, et al., 2013) have proposed and tried to evaluate the most effective classification approach for detecting Type 2 Diabetes Mellitus (T2DM) in rural adults, in this research the research team proposed an artificial neural network (ANN) and multivariate logistic regression (MLR) models to as a classifier tools for T2DM, the main intent of this research is to compare which the most accurate and reliable to use within ANN and

MLR, the research's result shown that ANN was 0.981 more accurate than the MLR models. In summarize, it can be concluded ANN as a part of computational intelligence approach provide more accurate result than regression process.

(Varma, Rao, Lakshmi, & Rao, 2014), proposed to develop a decision tree model to predict the occurrence of DM in someone, this research had been proposed in 2014. the research team convey the current traditional decision tree model has a problem with crisp boundaries, then they proposed to enhance the decision tree with fuzzy computation to prevent the sharp cut-off, this research used 336 of datasets and will be tested using MATLAB tools, the accuracy of this research is 75.8%, they convey that this result is still outperformed compared with the previous research and they have a plan to improve the accuracy in the future.

(Kandhasamy & Balamurali, 2015) research is comparing which algorithm that can provide the highest accuracy in predicting Diabetes Mellitus using data mining concept and techniques, hereinafter, some of algorithm that they want to compare are: Decision Tree, K-Nearest Neighbors, Random Forest and Support Vector Machines, and it shown that J48 decision tree generate the lowest accuracy than the other classifier.

In 2015, (Zhu, Xie, & Zheng, 2015) have proposed a dynamic weighted voting scheme which called multiple factors weighted. In this research, they focus on how multiple classifier systems (MCS) can perform in early detection of type 2 Diabetes Mellitus. Meanwhile, data sets also being used to measure and evaluate the accuracy of the proposed method, in summarize they have a plan to adopt genetic information to establish stronger output for the framework.

(Lukmanto & Irwansyah, 2015) have tried to adopt Fuzzy Hierarchical Model as a generated method of combining Fuzzy System and Analytic Hierarchy Process (AHP). In the same year, Feng Li & Kuo also proposed Hierarchical Fuzzy Classification. Extreme Machine Learning concept to classify PID dataset has been proposed by Ding et al in 2015. All

research proposed have not adopted either pre-processing or feature selection step.

Less than 4 years after (Beloufa & Chikh, 2013) proposed their powerful ABC algorithm to diagnose Diabetes Mellitus, in 2016 with the same datasets which is Pima Indian Diabetes Datasets, (Hayashi & Yukita, 2016) convey that currently lot of diagnostic methods for Diabetes Mellitus are black-box models, which is cannot provide the accurate reason underlying detection summarization to physicians. However, scientist need to think on how to generate a powerful and sustainable tool to diagnose Diabetes Mellitus to make physicians easier to understand the generated output by the system from medical standpoint. Therefore, (Hayashi & Yukita, 2016) proposed a rule extraction algorithm called sampling Re-RX with J48graft which combined with feature selection technique. Anyway, (Hayashi & Yukita, 2016) still declare that this diagnosis method still facing a remains complex problem which should be tested on more recent and complete diabetes datasets to ensure the accuracy.

In 2017, (Zheng, et al., 2017) have learned on how to diagnose Type 2 Diabetes Mellitus from the previous research, they propose a machine learning based frameworks to identify Type 2 Diabetes Mellitus through electronic health records. In summarize, they propose a data informed framework to identify subject based on the potential in Diabetes Mellitus via feature selection and machine learning, in this research they tried to evaluate various of machine learning framework such as Naïve Bayes, Decision Tree and Support Vector Machine by using sample data of generated electronic health records.

Also in the same year, (Kumar, Sharmila, & Singh, 2017) propose a SVM based approach to predict the most discriminatory gene target for Type 2 Diabetes Mellitus. The research team leverage Support Vector Machine classifier as a feature selector engine to discriminate the gene that has a potential in Type 2 Diabetes Mellitus.

(Barkana, Saricicek, & Yildirim, 2017) have proposed a research that related to performance analysis of descriptive statistical features to indicate retinal vessel segmentation caused by Diabetes Mellitus complication called diabetic retinopathy. This research will be evaluated by using Fuzzy Logic, Artificial Neural Network (ANN) classifier and Support Vector Machine (SVM). Barkana and team confirm that all classifier achieved the expected classification accuracies, they also state that they have a plan to enhance their proposed research to increase the current result in the future.

There are 5 key opportunities from the related works to that can be used as a baseline ensure we can drive this research to go through into correct process which enhancing the previous works, here are some of the key opportunities: Computational Intelligence approach leading the most usable technique and high accuracy result in detecting Diabetes Mellitus, Fuzzy Logic as a part of computational intelligence implementation method has a high frequency that been used as a classifier system in detecting Diabetes Mellitus, Nowadays, feature selection is a new trend phenomena to perform the execution of data preprocessing to increase the output accuracy prior the main computation process, There are lot of reference based on the related works that indicate there is an opportunity to leverage Support Vector Machine as a classifier method with promising to high accuracy output and Currently, more-less 17 years old flutter Pima Indian Diabetes (PID) datasets that was donated since 1990 are still exist and can be crowned as a quite powerful dataset which can be still usable to be leveraged in diagnosing Diabetes Mellitus.

2.2. Diabetes Mellitus

One of the most common characteristic of Diabetes Mellitus is when patient has a high blood glucose level caused by body deficiency or resistance to insulin which can lead to serious health complication (Beloufa & Chikh, 2013) (Varma, Rao, Lakshmi, & Rao, 2014). Meanwhile, as mentioned on previous statement that DM is a complex disease that can lead

to serious health complications, such as: Coronary Artery disease, Peripheral Vascular disease, Peripheral Neuropathy, Liver dysfunction and Sexual dysfunction in woman (Feinglos & Bethel, 2008).

According to (ADA, 2010) and (IDF, Diabetes Atlas 7th Edition, 2015) Diabetes Mellitus can be classified into 2 types. Below points will describe the 3 different types of Diabetes Mellitus based on the factors that cause the disease to reside in someone:

1. *Type 1 Diabetes Mellitus*, caused by the negative reaction from the body immune system that attacking *β-cell* that produced inside the pancreas, then impacting to number of insulin that produced by the body is not sufficient to control the blood glucose level. Unfortunately, the main cause of Type 1 Diabetes Mellitus has not been identified yet, meanwhile physicians believe that this disease caused by abnormal birth defect due there are lot of case happen in toddler.
2. *Type 2 Diabetes Mellitus*, can be caused by the character of the body that resistance to insulin, nor the number of produced insulin supplied by pancreas to keep control the body blood glucose level is not sufficient. There are some of mandatory points that need to be awarded due it can be impacting some on to be infected by Diabetes Mellitus: hereditary factor, maternity woman with the weight of the baby more than 9 lbs, someone with obesity, less physical activities and hypertension.

2.3. Feature Selection

Feature selection has been proven as an effective and efficient way to perform data preprocessing execution for high-dimensional data for data mining and machine learning (Li, et al., 2016). Feature selection, as a dimensionality reduction technique, aims to choose a small subset of the relevant features from the original ones by removing irrelevant, redundant, or noisy features. In summarize, Feature Selection or also known as attribute weighting, dimension reduction and so on, is a process in which attribute in

a data set has been reduced to a few attributes that really matter (Kotu & Deshpande, 2014).

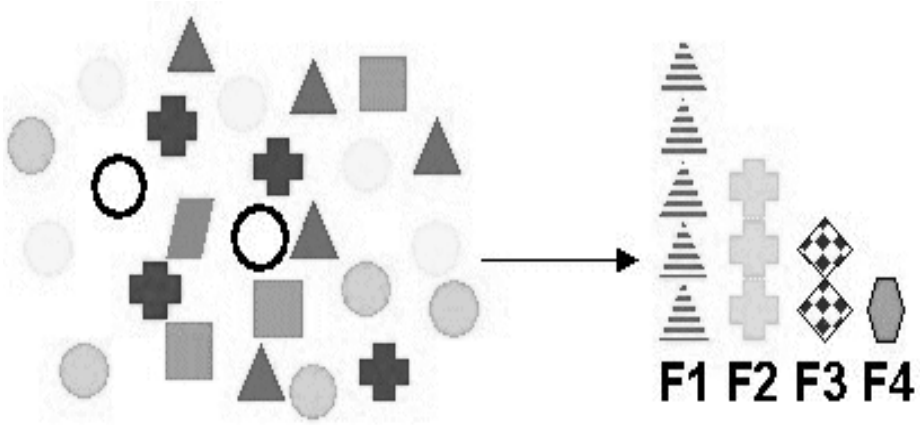


Figure 2.1 Illustration of Feature Selection technique

Based on previous statement, in this research the main consideration of leveraging feature selection is to find an optimal subset of the original features that sufficient to preserve good classification performance then the generated output accuracy might be increased.

2.4. F-Score Feature Selection

F-Score is one of powerful feature selection technique and served as well proved feature selection method that had being used in various of research which measures the discrimination of two sets of real numbers (Chen & & Lin, 2006) (Zemmoudj, Kemmouche, & Chibani, 2014). F-Score firstly introduced in 1979 as a weighted one-dimensional indicator of the two and was defined as their weighted harmonic mean (Song, Jiang, & Liu, 2017),

$$FScore_{\beta} \equiv (1 + \beta^2)\frac{\rho\pi}{\beta^2\pi + \rho} \qquad (1)$$

F-score expression was formed as a feature selection criterion, given training vectors $x_k, k = 1,.,m,$ If the number of positive and negative instance are n_+ and n_-, hereinafter, its value of ith feature is defined as:

$$\mathrm{F}\,(i) \equiv \frac{(x_i^{(+)} - x_i)^2 + (x_i^{(-)} - x_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{+} - x_i^{+})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_+}(x_{k,i}^{-} - x_i^{-})^2} \qquad (2)$$

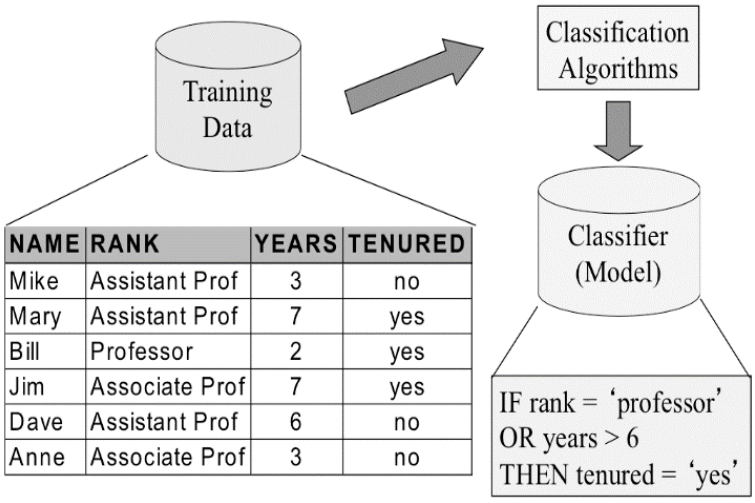
Where $x_i, x_i^{(+)}, x_i^{(-)}$ are the average of the ith feature of the whole, positive and negative data sets, respectively; $x_{k,i}^{+}$ is the ith feature of the kth positive instance and $x_{k,i}^{(-)}$is the ith feature of the kth negative instance (Chen & & Lin, 2006). The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative. Therefore, we leverage this score as a feature selection criterion.

2.5. Classifier Method

Classifier method or also known as Classification process has been built to be leveraged in every prediction task in a set of data class to find an exact identification from a data based on classification rules that had been founded from labeled sample data or supervised learning, then the un-recognized relevant data could be identified and labeled (Han, Pei, & Kamber, 2011), there 2 steps of process in Classification:

1. Learning Step

   Learning step is a phase when the training or sample data will be analyzed by the classification algorithm, then the result of the learning will be represented in the form of classification rules.

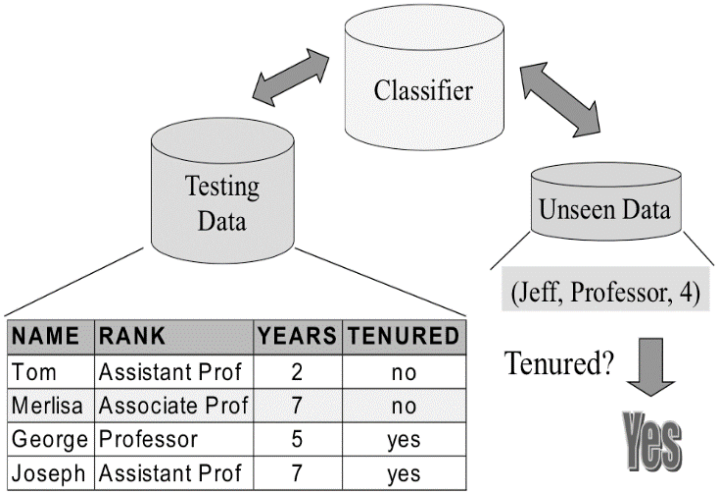


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Figure 2.2 Learning step of classification process

2. Classification Step

   Classification step is a phase where the generated classification rules from learning step will be used to predict the new un-recognized data. In this step, the accuracy test also going to be executed by using the test data to measure and get the accuracy value of classification rules that had been represented.



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Figure 2.3 Classification step of classification process

2.6. SVM Classifier

Support Vector Machine (SVM) is a discriminative classifier method that had been proven and well-known classifier method that eligible and powerful to be adopted in various of research area that need to perform classification execution in the process (Chen & & Lin, 2006) , formally SVM was defined by a separating hyperplane.

Basically, Classification is one of the main areas of application for Support Vector Machine (Steinwart, 2008), SVM classifier method will train the labeled training data or also known as supervised learning to recognize the relevant input data as describe in above Classifier Method sub-section, then the generated result of each labeled data will be classified based on their discrimination states that will be separated by the hyperplane. As an example, if the goal of leveraging Support Vector Machine classifier

algorithm in the research is to make a diagnosis of the disease, then the goal is to estimate a response that only has two states (Steinwart, 2008),

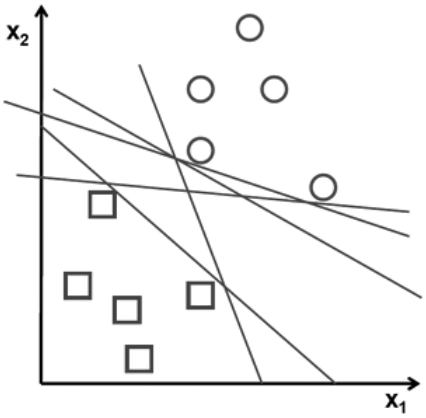


Figure 2.4 Straight Line to Separate Two Classes

Figure 2.4 illustrate how multiple straight line can cut-off the differences and offer a solution to solve the classification problem by separating and classify each of un-relevant nodes. Meanwhile, to ensure the SVM classifier algorithm work properly, frankly this is un-acceptable or may have a potential to reduce the classification accuracy when the straight line get too close with the nodes because it will be noise sensitive and it will not generalize correctly.

Therefore, the main intent of SVM algorithm is to find the line which passing as far as possible from all sides of the nodes and generating the maximum margin from each side, this chosen straight line also called Optimal Hyperplane.

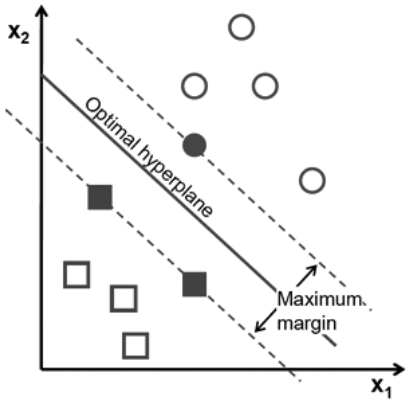


Figure 2.5 Optimal Hyperplane for Maximum Margin

Based on Figure 2.5 it was cleared that the focus and intent of leveraging SVM classifier algorithm is to get the most optimal Hyperplane to optimize the classification accuracy.

2.7. Fuzzy Logic

Fuzzy Logic as a part of Computational Intelligence had been introduced firstly by Lotfi A. Zadeh in 1965, is a knowledge based system or also known as a rule-based system, which established to avoid sharp-cut off computation execution caused by crisp set, then the Fuzzy logic will convert the crisp sets into fuzzy sets (Ross, 2010:34). In summarize, the main intent of leveraging Fuzzy Logic is to optimize the classifier computation process by avoiding the sharp cut-off performing by crisp or classical classification method.

In Fuzzy Logic, we need to define the membership degree of each fuzzy sets which called membership function. In Fuzzy Logic implementation, the Membership Function will define blurring level or also known as fuzziness of each elements in Fuzzy sets (Ross, 2009). At this point, it can be concluded the main difference between crisp sets computation and fuzzy sets computation provided by Fuzzy logic. As shown in Figure 2.6, crisp sets computation only has two kinds of membership degree which is 0 and 1, hereinafter, the sharp cut-off will be implemented in crisp computation and there is a potential to reduce the output accuracy.

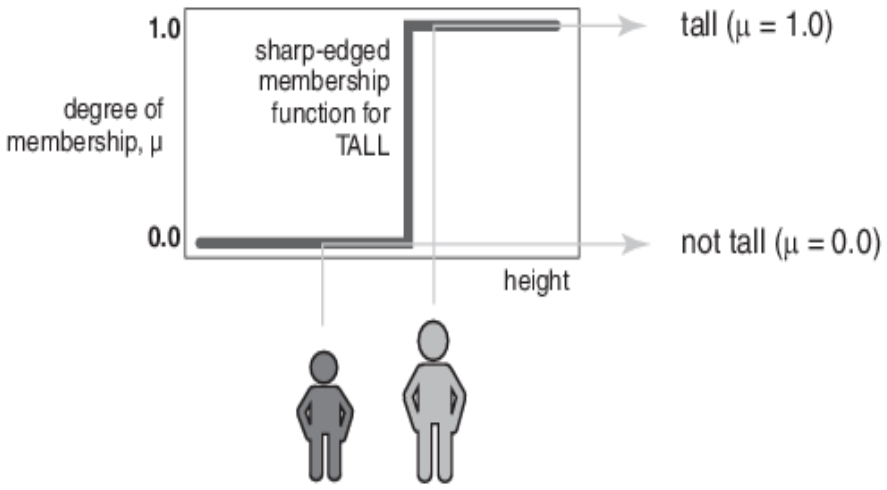


Figure 2.6 *Membership functions of crisp sets*

Meanwhile, in Fuzzy sets computation as shown in Figure 2.7 the membership degree will be formed a continuous edge which allow the elements in fuzzy sets possible to generate the highest possibility or membership degree in each of area of membership. In summarize, Fuzzy sets will establish the membership degree from 0 to 1.

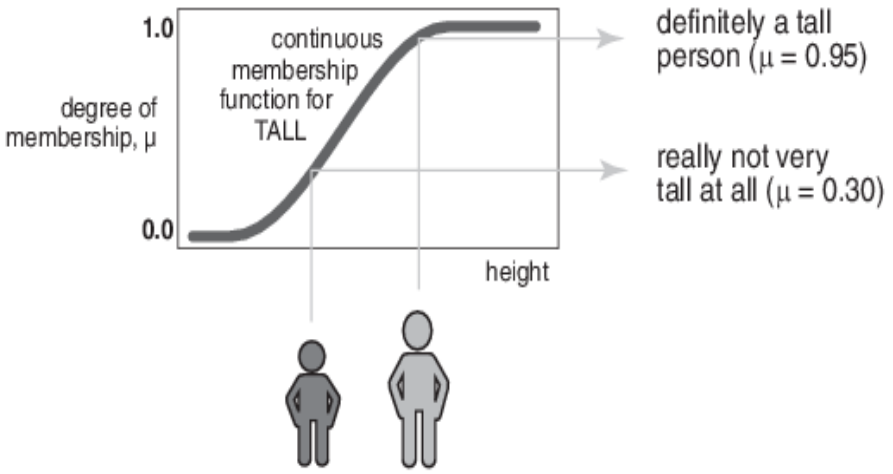


Figure 2.7 *Membership functions of fuzzy sets.*

2.8. Fuzzy SVM Classifier

Combining Support Vector Machine (SVM) classifier and Fuzzy modelling has been proposed by (Ramanathan & Sharma, 2015). The main objective is to drive better accuracy in classification process, especially in area of computational intelligence techniques that have been developed for classification process in recent years. In this technique, Fuzzy systems has adopted to classify the data while the SVM classifier will be used to generate the fuzzy rules.

In general, the inputs are fuzzified into fuzzy sets using triangular and trapezoidal membership function that driven by designed rules generated by SVM classification result which trained using Pima Indian Diabetes Dataset.