Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

11th International Conference
on Electrical Engineering
ICEENG 2018

Novel Frequency Domain Classification Algorithm Based On Parameter Weight Factor Computation

Ahmed M. Salah EL-Bohy*, Atallah I. Hashad*, Hussien Saad Taha** and Ahmed Azouz**

ABSTRACT

This paper considered a state of art as it employs a brand new technique to estimate the missing values in the dataset helping the classifiers to classify the data with better accuracy. also determining the effect of each attribute on the accuracy that enables researchers to get better or same accuracy with less number of attributes saving processing time, RAM, and memory needed. The aim of this paper is to improve accuracy of expecting Hepatitis mortality using worldwide dataset from Ljubljana University. We present an implementation of two brand new classification techniques. Using confusion matrix and K-fold cross validation technique to calculate classification accuracy. Two experiments have been done and the experimental results show that using the correlation in frequency domain after computing the weight factor for each attribute achieved better accuracy than using the subtraction method in time domain.

KEYWORDS

Hepatitis, K-fold cross validation, Classification, Classification in frequency domain Confusion Matrix.

I. INTRODUCTION

Some diseases like hepatitis has a very difficult diagnosis task for a doctor, where doctors usually determine decision by comparing the current test results of patients with another one who has the same condition. Hepatitis is one of the most common diseases all around the world especially in Egypt; as it represents 22% of hepatitis cases around the world. This encourages us for proposing new methods to improve the outcomes of existing methodologies, as well as helping doctors and specialists to diagnose hepatitis disease survival [1].

* **Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt**
** **Egyptian Armed Forces**

Hepat (is a Greek word means) 'liver' and suffix–***its*** denotes 'inflammation' of the liver and may be due to infectious or non-infectious causes. The five types of hepatitis viruses are Common infectious causes of liver inflammation and some like Hepatitis A (HAV), B (HBV) and C (HCV) are more frequently seen infectious agents. Inflammation may lead to death of the liver cells (hepatocytes) which severely compromises normal liver function. Acute HBV Infection (less than 6 months) may resemble the fever, flu, muscle aches, joint pains and generally being unwell. Symptoms denote that situations are dark urine, loss of appetite, nausea, vomiting, jaundice, pain up the liver. Chronic hepatitis B is an infection persisting more than 6 months, the clinical features of that state correspond to liver dysfunction, so signs like these may be noticed: enlarged liver, splenomegaly, hepatosplenomegaly, jaundice, weakness, abdominal pain, confusion and abdominal swelling [2]. The success of treatment depends on an early recognition of the virus, which achieves more exact and less violent treatment options and mortality from Hepatitis falls.

Recently, data-mining has become one of the most treasured tools for handling data in order to create valuable information for decision-making [3]. Supervised learning, including classification is one of the most significant brands in data mining, with a recognized output variable in the dataset. Classification methods can achieve high accuracy in classifying mortality cases. Several papers about applying machine learning procedures for survivability analysis in the field of Hepatitis diagnostic. Here are some examples:

Using Support Vector Machines and Wrapper Method for predicting Hepatitis was introduced achieving maximum accuracy of (74.55%). [4]. but we note that applying SVM classifier only get higher accuracy than the mentioned accuracies even with feature selection as it achieves the accuracy of (79.38%) [5,6].

Improving the accuracy of SVM algorithm using feature selection [7]. Using SVM with Chi-Square achieved accuracy of (83.12 %), but we note that applying another classifier (Logistic, Simple Logistic, SMO, RF, J48) gets higher accuracy than the mentioned one as it comes with the accuracy of (85.17%) [5,6].

Here in this paper we will introduce a new technique computing the weight factor and having the most effective (7) attributes to compare our result with that one stated in the paper named prediction of hepatitis prognosis using support vector machines and wrapper method [4]. Where the enormous note is that paper stated that the results are calculated with (7) attributes out of (25) ones although the original number of attributes in the dataset is 20 only the achieved accuracy is (85%) Vs. (90.4%) proposed accuracy.

The rest of this paper is prearranged like this: In sector II, Classification algorithms are discussed. In sector III Evaluation principles and proposed methodologies are discussed. In sector IV reports the experimental results. Finally, Sector V introduces the conclusion and future work.

II. CLASSIFICATION ALGORITHMS

A Bayesian belief network is sometimes named a Bayes net, a belief net, or a causal network; it is a directed, acyclic graph, indicating conditional dependencies. It can be used to guess the probability of events. The Bayesian decision rule assurances minimum error if likelihoods and prior probabilities are known [9].

Decision Tree (DT) Tree that the root and each interior node is marked with a question. Can be used without the computer and are fairly easy to understand. Positions of attributes in the tree, especially the top ones, often directly correspond to the domain expert's knowledge. However, in order to produce general rules, these methods use pruning, which drastically reduces the tree sizes. Correspondingly, the paths from the root to the leaves are shorter, containing only few, although most informative attributes. In many cases the physicians feel that such a tree describes very poorly the diagnoses and is therefore not sufficiently informative However, as mentioned earlier, the structure of generated trees by Assistant-R is more human-like, which was confirmed in several diagnostic tasks The arcs represent each possible answer to the concomitant question. Each leaf node represents a forecast of a problem solution. A prevalent technique for classification; Leaf node leads the class to which the corresponding tuple belongs. Its model is a computational model comprising of the three parts: Decision Tree Algorithm to create the tree Algorithm that applies the tree to the data, creation of the tree is the most exciting part. Processing is mostly a search similar to that in a binary search tree (although DT may not be binary). Advantages: Easy to understand, easy to generate rules [10].

Support Vector Machine (SVM) SVMs are amongst the best (and many believe are definitely the best) “on-the-shelf” supervised learning algorithm. It is derived from statistics in 1992.SVM is widely used in multiple applications pattern recognition, classification and regression. The SVMs work on an underlying principle, which is to insert a hyper-plane between the classes and orient it in such a way to keep it at the maximum distance from the nearest data points these data points, which appear closest to the hyper-plane, are known as Support Vectors [8], [11].

Sequential Minimal Optimization (SMO) humble, easy to implement, is generally quicker, and has better scaling properties for difficult SVM problems than the usual SVM training algorithm SMO quickly solve the SVM QP problem without extra matrix storage (the memory used is linear with the training dataset size) and without using numerical QP optimization steps at all. SMO can be used for online learning. While SMO has been shown to be operative on sparse datasets and especially fast for linear SVMs, the algorithm can be extremely slow on non-sparse datasets and on problems that have many support vectors. Regression problems are especially prone to these matters because the inputs are usually non-sparse real numbers (as opposed to binary inputs) with solutions that have many support vectors. Because of these restrictions, there have been limited reports of SMO being successfully used on regression problems. [12]

Logistic Regression (LR) is a famous well-known classifier; it could be used to extend classification results into a deeper analysis. It is not widely used due to its slow response in data mining especially when compared with SVM in large datasets (not our case) it models the relationship between a dependent and one or more independent variables, and consents us to look at the fit of the model as well as at the significance of the relationships [13]

Simple Logistic: We use simple logistic regression when we have one nominal variable with two values (dead/alive, male/female) and one measurement variable. The nominal variable is the dependent variable, but the measurement variable is not it is an independent one. Simple logistic regression is analogous to linear regression, except the dependent variable is nominal, not a measurement. [14]

Random Forest: Random forests change how the classification or regression trees are constructed. In standard trees, it uses the best split among all variables to split each node In a random forest, each node is split using the best among a subset of predictors arbitrarily chosen at that node. it works by one of two methods, boosting and bagging.it has the advantages of: handling thousands of input variables without deleting any variable, giving estimation of variables importance in the classification, and It also has an active method for estimating missing data and keeps accuracy when a large amount of the data are missing [15].

III. EVALUATION PRINCIPLES AND PROPOSED METHODOLOGIES

Two methodologies applied to the Ljubljana dataset [16], in both methods data preprocessing including data cleansing, removing unwanted parameters and normalization are common procedures done before applying any classification algorithm.

A comparison has been made with WEKA results [5,6] where WEKA, formally called Waikato Environment for Knowledge Learning (also, the WEKA is a flightless bird found only in the islands of New Zealand), is a computer program that was developed at the University of Waikato in New Zealand for identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic principle of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the public license agreement. WEKA application has been written in C and rewritten in Java. It is a user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, like numeric values. The WEKA application allows beginner users to use options and visual interfaces simply. WEKA has workflow support via its Knowledge Flow utility. Finally, WEKA is the unique tool which has a built in multi classifier fusion [17].

Method # 1:

Checking the data for presence of missing values if so, comparing the instance that has missing value(s) with all other instances then take summation of the differences and taking the minimum value of the absolute summation (Min. Difference) then placing the missing value of that instance as the most like instance.

Until having no missing values in the whole instances of the dataset these procedures are applied.

Specifying the number of iterations and concerning Mont Carlo simulation to have the error RMS getting the accuracy percentage.

Splitting the dataset randomly into two separate datasets training and testing datasets (based on K fold method).Based on the same idea of subtracting the absolute sum of the test instance from all training dataset instances excluding the decision attribute. Selecting the minimum and consider it as a decision like instance, repeating that stated procedure by increasing the initial condition counter of the

testing data until getting the end of the testing dataset. Evaluating the accuracy of the proposed method based on comparing the decision with the pre-known decision (Supervised Learning) then uses the confusion matrix as an evaluation technique. The confusion matrix is an imagining implement usually used to show presentations of classifiers. It is used to illustrate the relationships between real class attributes and predicted ones. The grade of efficiency of the classification task can be computed with the number of exact and unseemly classifications in each conceivable value of the variables being classified in the confusion matrix [18].

Table 1. Confusion matrix

		Predicted Class	
		Negative	Positive
Outcomes	Negative	Tn	Fn
	Positive	Fp	Tp

For instance, in a 2-class classification problem with two predefined classes (e.g., Positive, negative) the classified test cases are divided into four categories:

- True positives (Tp) correctly classified as positive instances.
- True negatives (Tn) correctly classified negative instances.
- False positives (Fp) incorrectly classified negative instances.
- False negatives (Fn) incorrectly classified positive instances.

To evaluate the classifier performance, we define accuracy term that is defined as the entire number of truly classified instances divided by the entire number of available instances for an assumed operational point of a classifier.

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad \text{“1”}$$

The whole procedure of that method is shown in Fig.1.
where: N: Number of instances that have missing value(s), n: Counter for instances, Num_I: Number of iterations. LLL: Number of instance in test dataset, a: Current instance under test, and I: Iteration counter.

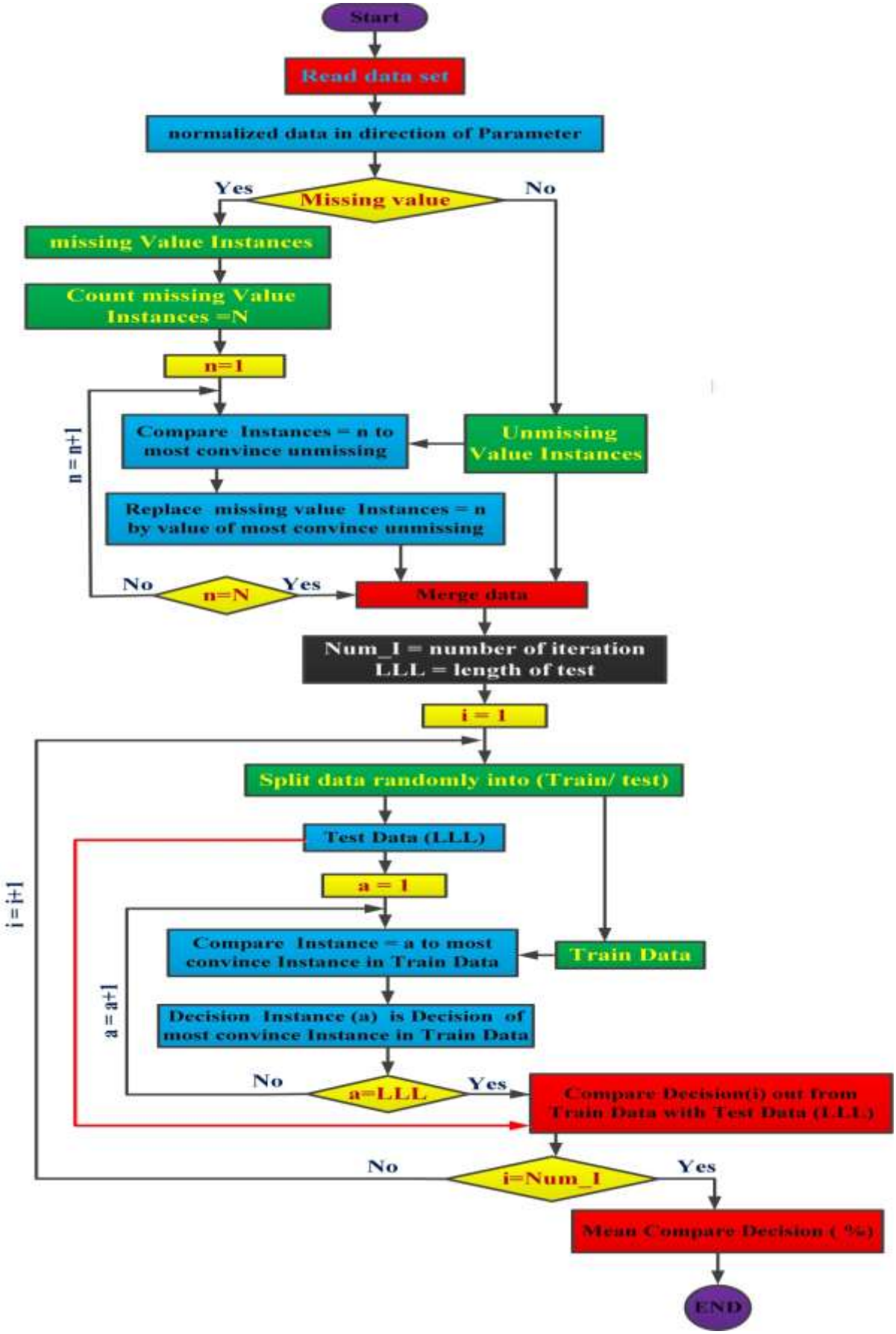

Fig.1. Method # 1 Difference in time domain

Method # 2:

Consuming time correlating signals in time domain to shift the signal by the shift (2n-1) where n is the signal length, and that means the need to apply this procedure twice (in row and column direction). The first time correlating (in column direction) the decision of (140) instance having (n=19) getting a weight for each attribute and determine its' effect on the decision. Moreover, the second time correlating (in row direction) the set of (19) attributes of the test dataset with the same (19) attributes of the training data set having (n=140) to get the most similar decision to the (140) instances was the motive to execute the correlation in frequency domain.

Starting our iteration by giving the loop initial condition then Determine the missing value(s) in each instance by correlating (Frequency Domain) the instance with all completed (non-missing) instances taking the absolute value of the result then summing it to have the maximum correlated instance and copying its' attribute at the missing attribute position of that instance. as judging a complete non missing instance not only more easy process but also higher accuracies are guaranteed.

Split the dataset randomly into training and test datasets then compute the weight factor of each attribute of the training dataset separately by correlating the decision attribute (in frequency domain) with each attribute of the remaining 19 attributes. Summing the absolute values of the correlation process then sort the results in descending order and use only the first seven attributes (7 out of 19) which is the same number of attributes taken in the Paper named prediction of hepatitis prognosis using support vector machines and wrapper method of the literature survey[4].

Applying Fast Fourier Transform (FFT) to the test dataset then taking the conjugate to multiply it by the training dataset FFT then get the IFFT having the absolute value and the summation of the cross correlation as shown in Fig.2 below. Then sorting the results in descending order and use the voting methodology between the highest three to get a decision by at least 2 out of 3 as the live / die decision will be taken as live when two or more of the three cases class are live and will be taken as die when two or more of the three cases class are die

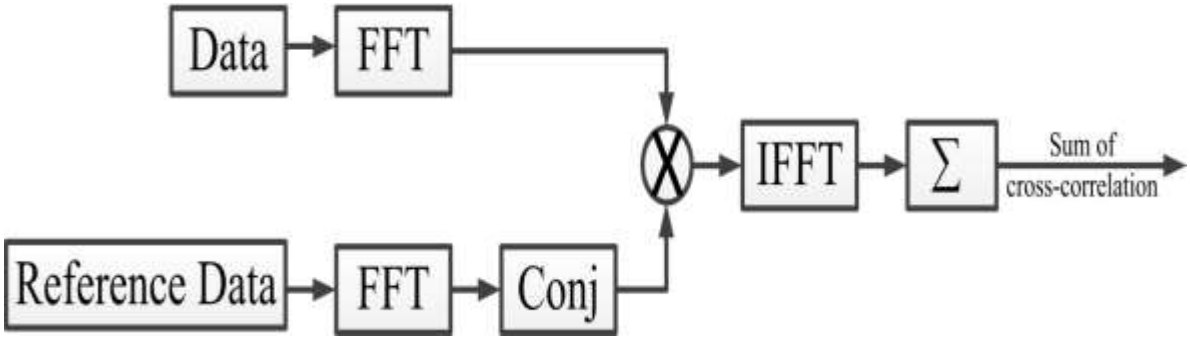

Fig.2. Frequency domain transformation

Using the confusion matrix mentioned above in the first method and illustrated in table.1 to calculate the classifier accuracy percentage.

The frequency domain correlation method is illustrated in Fig.3.below where: L_Test: Length of test dataset and N_I: Iteration counter.

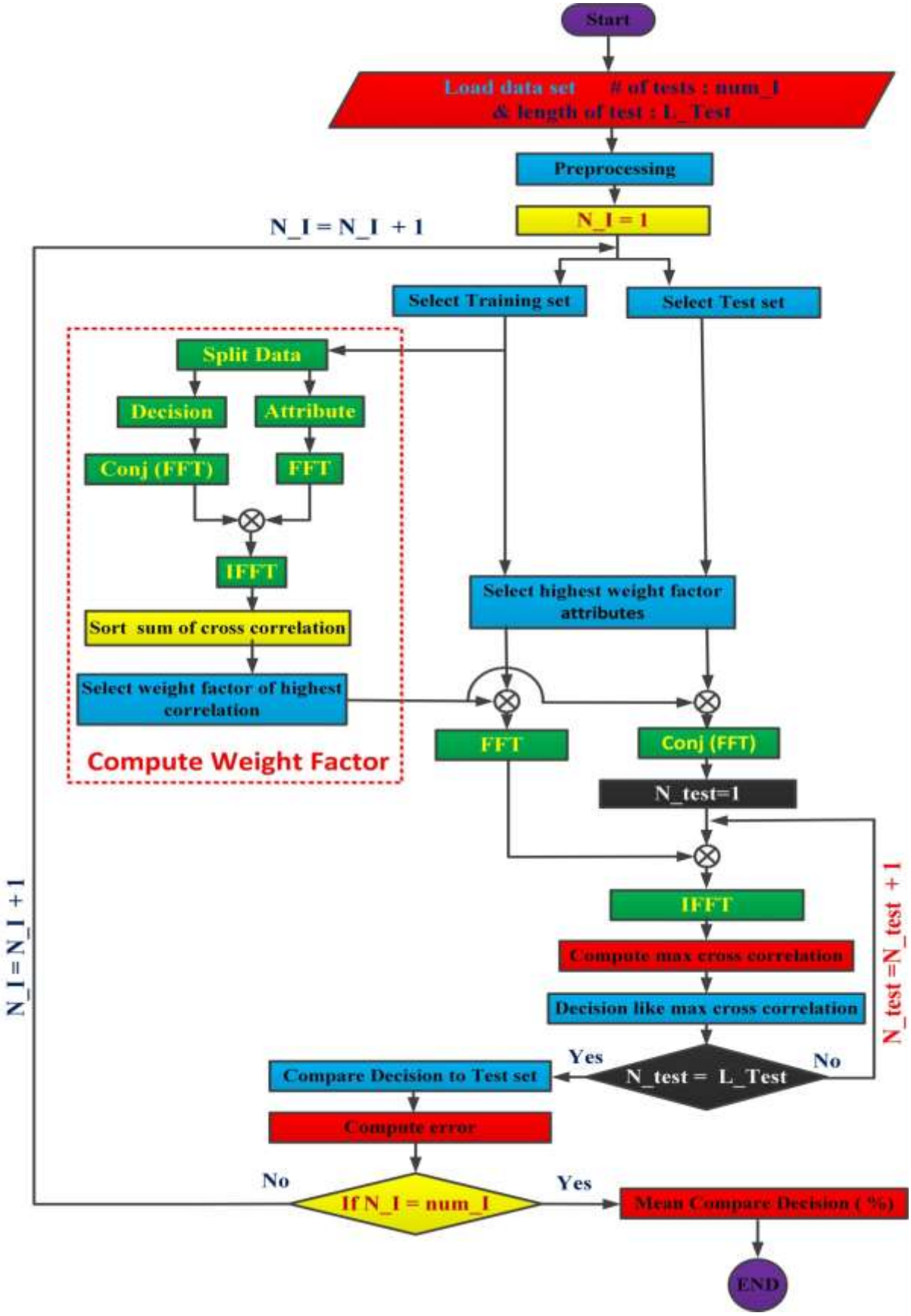

Fig.3. Method # 2 Frequency domain classification

IV. EXPERIMENTAL RESULTS

Both methods has been applied to the complete dataset the whole 155 instances. One thousand iterations has been applied for both experiments guarantees that all instances have been included in both methods, a split of dataset randomly into training & test (20 instances as a test & 135 instances as training) in each iteration.

The next histograms show that all instances have been selected more than 100 times. As shown in Fig.4 for the first method difference in time domain and Fig.5 for the second method cross correlation in frequency domain

Accuracy has been calculated using equation “1” above depending on the confusion matrix concept explained above in sector III of this paper.

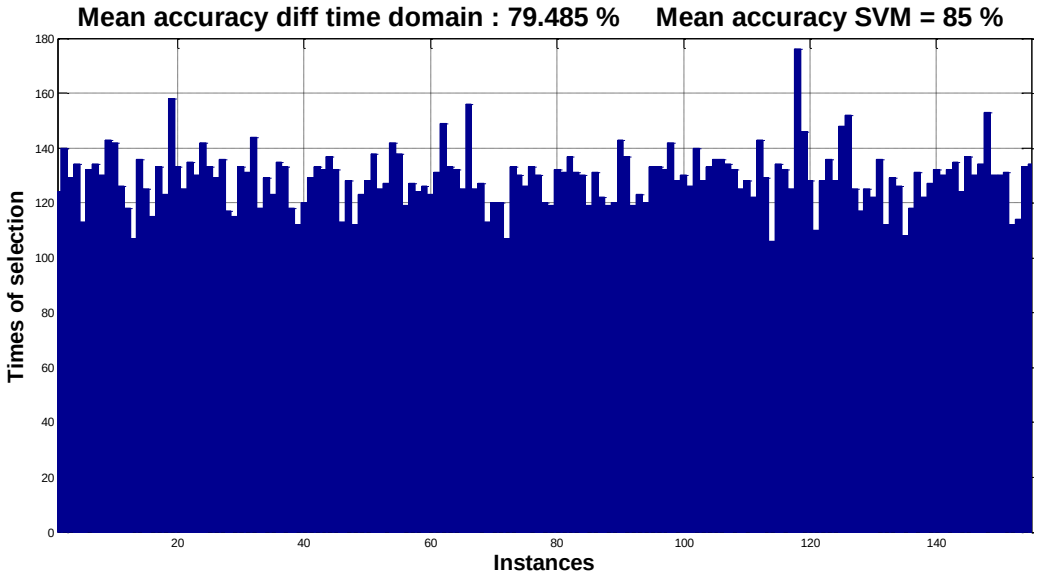

Fig.4 Histogram for (first) time domain method

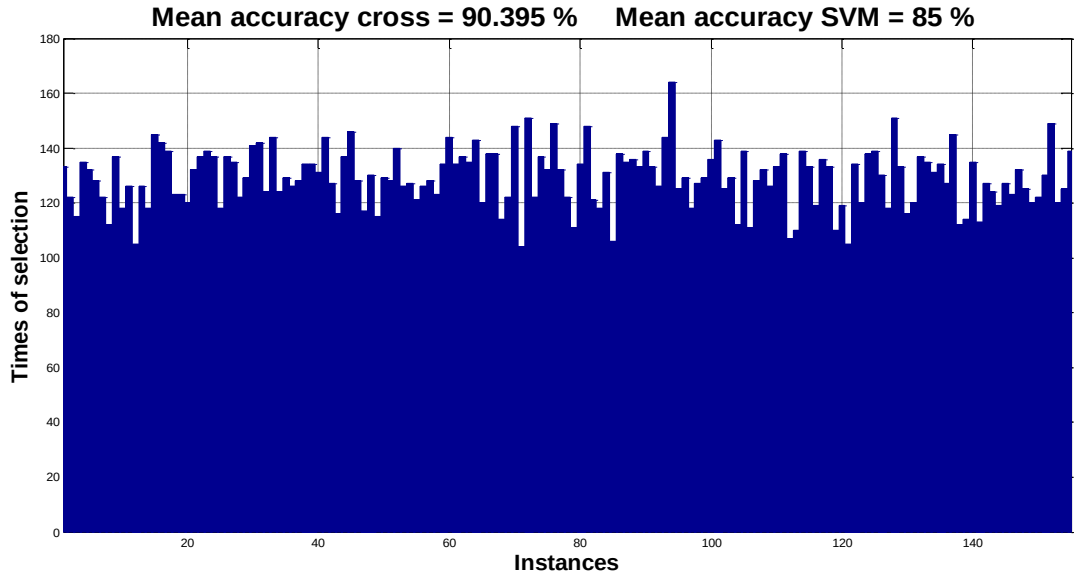

Fig.5 Histogram for (second) frequency domain method

Table 2 Experimental Results

Software	Algorithm	Accuracy (%)	Proposed Algorithm	Notes
WEKA (Ready Made)	Bayes Net	83.21	Nil	references [5,6]
	SVM	79.38		
	Logistic	82.58		
	SGD	84.54		
	Simple Logistic	83.88		
	SMO	85.17		
	K*	81.96		
	J48	83.79		
	RF	85.13		
MATLAB	Time Domain	SVM: 85% (Pre-Programmed)	79.40%	
	Frequency Domain		90.4%	

As illustrated in table 2 above: Proposed MATLAB Algorithm in frequency domain achieves the best accuracy (90.4%) while the time domain proposed algorithm achieves a slightly higher (almost the same) accuracy of the WEKA SVM algorithm.

Reached accuracy in frequency domain is higher than all accuracies in time domain also it is higher than the accuracies proposed in [5], [6] using classifiers fusion.

Accuracy can be compared with the accuracy gained from classifiers fusion in case of reduced or no missing values in the dataset, denoting perfect determination for the missing values.

V. CONCLUSION AND FUTURE WORK

Using frequency domain achieving better accuracy, besides we got the advantages of reducing processing time, RAM, and internal storage. also achieving higher accuracy with less attributes by determining the weight factor for each attribute helps doctors to specify the most important medical examinations helps doctors to diagnose and take decisions saving more and more patients , less over heads and cost for patient.

Fusion between different classifiers in frequency will be the future work.

REFERENCES

[1] World Health Organization at http://www.who.int

[2] Tomasz KANIK. “Hepatitis disease diagnosis using Rough Set modification of the pre-processing algorithm.” Information and communication technologies international conference (2012).

[3] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth."From data mining to knowledge discovery in databases." AI magazine 17.3 (1996).

[4] A.H.Roslina and A.Noraziah. "Prediction Of Hepatitis Prognosis Using Support Vector Machines And Wrapper Method." Seventh International Conference on Fuzzy Systems and Knowledge Discovery (2010).

[5] Ahmed M. Salah EL-Bohy, Atallah I. Hashad, and Hussien Saad Taha "Performance Evaluation Of Hepatitis Diagnosis Using Single and Multi-Classifiers" International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 04, April 2015.

[6] Ahmed M. Salah EL-Bohy, Atallah I. Hashad, and Hussien Saad Taha "Highly Reliable Hepatitis diagnosis with multi classifiers" ASAT 16, 2015, Cairo, Egypt.

[7] Varun Kumar.M , Vijay Sharathi.V and Gayathri Devi.B.R. " Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy." International Journal of Computer Applications (0975 – 8887) Volume 51– No.19, August 2012.

[8] C. Barath Kumar , M. Varun Kumar , T. Gayathri, and S. Rajesh Kumar " Analysis and Prediction of Hepatitis Using Support Vector Machine." International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2235-2237.

[9] Jyotirmay Gadewadikar, Ognjen Kuljaca1, Kwabena Agyepong, Erol Sarigul3, Yufeng Zheng and Ping Zhang, Exploring Bayesian networks for medical decision support in breast cancer detection, African Journal of Mathematics and Computer Science Research Vol. 3(10), pp. 225-231, October 2010.

[10] Ross Quinlan, (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

[11] Chen Y., Wang G., and Dong S. "Learning with progressive transductive support vector machine", Pattern Recognition Letters, Vol. 24, Pages: 1845-1855

[12] John C. Platt. "Sequential Minimal Optimization:A Fast Algorithm for Training Support Vector Machines." Technical Report MSR-TR-98-14April.

[13] Mitchell T. "Machine Learning, McGraw Hill." (1997).

[14] John H. McDonald. "Handbook of Biological Statistics." at http://www.strath.ac.uk

[15] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[16] UCI Machine Learning liver, http://archive.ics.uci.edu/ml/datasets.html.

[17] Ralf Mikut and Markus Reischl "Data mining tools", 2011 John Wiley & Sons, Inc. WIREs Data Mining Knowledge Discovery 00 1–13 DOI: 10.1002/widm.4, pp. 1-13, (2011).

[18] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees and A.Zanasi, "Discovering data mining from concept to implementation". UpperSaddle River, N.J.: Prentice Hall, 1998.