Circuits and Systems, 2016, 7, 701-708
Published Online May 2016 in SciRes. http://www.scirp.org/journal/cs
http://dx.doi.org/10.4236/cs.2016.76059

Multitude Classifier Using Rough Set Jelinek Mercer Naïve Bayes for Disease Diagnosis

S. Prema[1], P. Umamaheswari[2]

[1]Information and Communication Engineering, Anna University, Chennai, India
[2]Computer Science and Technology, MIT Campus, Anna University, Chennai, India
Email: asokprema555@gmail.com, dr.umasundar@gmail.com

Received 29 February 2016; accepted 9 May 2016; published 12 May 2016

Abstract

Classification model has received great attention in any domain of research and also a reliable tool for medical disease diagnosis. The domain of classification model is used in disease diagnosis, disease prediction, bio informatics, crime prediction and so on. However, an efficient disease diagnosis model was compromised the disease prediction. In this paper, a Rough Set Rule-based Multitude Classifier (RS-RMC) is developed to improve the disease prediction rate and enhance the class accuracy of disease being diagnosed. The RS-RMC involves two steps. Initially, a Rough Set model is used for Feature Selection aiming at minimizing the execution time for obtaining the disease feature set. A Multitude Classifier model is presented in second step for detection of heart disease and for efficient classification. The Naïve Bayes Classifier algorithm is designed for efficient identification of classes to measure the relationship between disease features and improving disease prediction rate. Experimental analysis shows that RS-RMC is used to reduce the execution time for extracting the disease feature with minimum false positive rate compared to the state-of-the-art works.

Keywords

Classification Model, Disease Diagnosis, Rough Set Model, Feature Selection, Multitude Classifier, Mercer Naïve

1. Introduction

In a conventional classification model, the classification strategy identifies and selects the best classifier on the basis of experimental assessment with various individual classifiers. In a diversion from the conventional approach, the use of Multitude Classifier System (MCS) has been presented as an alternative approach to improve classification accuracy of the disease being detected.

How to cite this paper: Prema, S. and Umamaheswari, P. (2016) Multitude Classifier Using Rough Set Jelinek Mercer Naïve Bayes for Disease Diagnosis. *Circuits and Systems*, **7**, 701-708. http://dx.doi.org/10.4236/cs.2016.76059

Potential Management of Ventricular Arrhythmias (PPM-VA) [1] identified the stroke pattern and management of ventricular Arrhythmias that were largely related to stroke for the effective diagnosis of disease at an early stage. Prediction of Events using Spatio Spectro Temporal Data (PE-SSTD) [2] provided with a case study on stroke resulting in the accuracy of the disease being diagnosed. However, both the above methods lack the class accuracy of disease diagnosis with the increase in the feature.

To improve the detection rate of disease for Interstitial Lung Disease (ILD) in [3], a method was presented improving the detection of disease at an early stage. However, the classification accuracy remained unsolved. To address the issues related to class accuracy, Striatial Binding Ratio (SBR) was used in [4] to improve the class accuracy rate. A functional classification model [5] for early detection of heart failure using classification schema was presented for improving class accuracy rate. Classification methods for bipolar disorders [6] were designed to improve diagnostic reliability. Another method using neural network and decision based support system was designed [7] to improve the classification accuracy rate. However, accuracy with respect to scalability remained unaddressed.

An efficient classification approach using ANN and Feature Subset Selection [8] was presented to improve the accuracy of disease being detected. In [9], with the objective of improving classification rate, Diagnostic and Statistical Manual of Mental Disorders was presented.

In accordance with the above-mentioned advantages of both disease classification and disease diagnosis, in this paper, a new framework called Rough Set Rule-based Multitude Classifier (RS-RMC) is proposed to increase the disease prediction rate and efficiency of the classification accuracy of disease being diagnosed.

2. Design of Rough Set Rule-Based Multitude Classifier

To address the problem of disease diagnosis at an early stage, a framework is proposed based on Rough Set Rule-based Multitude Classifier. The Rough Set Rule-based Multitude Classifier uses disease features based on similar type of medical diseased data to identify the relationship between the disease features for efficient classification. **Figure 1** shows the block diagram of Rough Set Rule-based Multitude Classifier.

As shown in **Figure 1**, Cleveland Heart Disease dataset extracted from UCI repository is given as input. The block diagram shows a two-stage process. In the first stage, Rough Set Feature Selection model is applied to the input dataset to extract the disease features. Feature Reduct applied in Rough Set model reduce the disease feature without losing significant information. This is performed through lower and upper approximation without changing the values of disease features.

The second stage goes through the Multitude Classifier model called as the Mercer Naïve Bayes Classification model. This is performed using Naïve Bayes Disease Classifier algorithm aiming to improving the class accuracy. Finally, the Jelinik Mercer Multitude Classifier is used as a smoothing technique for obtaining approximation function increasing the disease prediction rate.

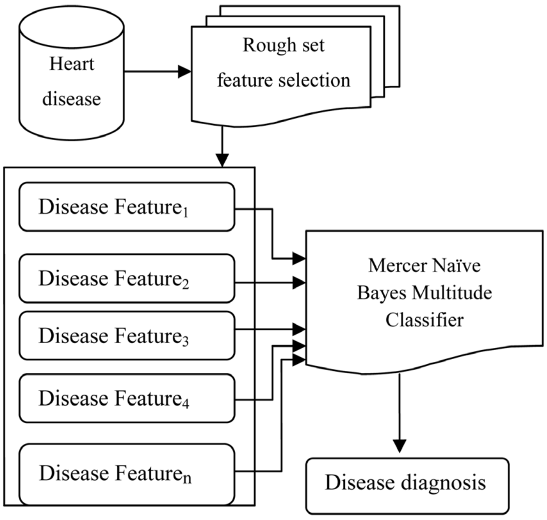


Figure 1. Block diagram of rough set rule-based multitude classifier.

2.1. Construction of Rough Set (Model) Feature Selection (Minimize Execution Time)

The first step in Rough Set Rule-based Multitude Classifier (RS-RMC) framework is the construction of Rough Set Feature Selection (RSFC) to reduce the complexity by minimizing the redundant disease features contained within the set of feature patterns. The feature selection using Rough Set model minimizes the redundant disease features by selecting those significant disease features that are most essential for Multitude Classifier represented in the pattern set.

Let us consider "*DFP*" the set of all disease feature patterns, "*F*" the set of all disease features, then the value of disease feature "*a*" in disease feature pattern "*P*" is given as below

$$f\left(a, P\right), \text{ where } a \in DFP \ \& \ P \in F \tag{1}$$

From (1), all disease features with disease feature patterns are identified. In order to reduce computational complexity and memory requirements, the disease feature patterns with the features identified is herein treated as a Rough Set model. The objective of using Rough Set model is to reduce the redundant features and therefore minimizes the execution time to obtain relevant features.

With the reduced features, the indiscernible relationship for identifying the features is obtained. For any two disease patterns, $\left(a, b\right)$ in "*DFP*" indiscernible relation for disease features is given as below

$$f\left(a_i, P\right) = f\left(b_i, P\right), \text{ for all } a \in P \tag{2}$$

For example, given the disease features (Chest Pain, Blood Pressure, Heart rate), the indiscernible set are, "$\left\{\left(P_1, P_3\right), \left(P_2, P_6\right), \left(P_3, P_5\right), \left(P_4\right)\right\}$". The Rough Set Feature Selection in RS-RMC framework performs two operations called, lower and upper approximation to measure the significance of the disease feature from the disease feature pattern set. Then, the lower and upper approximation is obtained as given below

$$DFP\left(A\right) = \left\{x \in DFP\right\} | \left[x_R\right] \subseteq A \tag{3}$$

$$DFP'\left(A\right) = \left\{x \in DFP\right\} | \left[x_R\right] \subseteq A \neq \varnothing \tag{4}$$

From (3) and (4), the disease features using lower approximation is identified, where "*A*" is the set of disease feature patterns in "*DFP*" that are surely in "*A*". On the other hand the upper approximation with "*A*" as the set of disease feature patterns in "*DFP*" is probably in "*A*". Finally, with the objective of reducing the disease feature set without losing significant information, disease feature Reduct Set is applied to the resultant set. A disease feature Reduct Set is then defined as a subset "*P*" of disease features (*i.e.*, reduct disease features) "*f*" is given as below,

$$\alpha f\left(DFP\right) = \alpha P\left(DFP\right) \tag{5}$$

By applying (5), the RS-RMC framework searches for disease feature reduct set of least cardinality. As a result, RSFC preprocesses disease feature patterns without changing the values of the disease features, aiming at minimizing the execution time for obtaining the disease features. Using this resultant disease features relationship between features helps in the easy and early diagnosis of disease.

2.2. Jelinik Mercer Naïve Bayes Classifier

The second step in Rough Set Rule-based Multitude Classifier (RS-RMC) framework is the effective identification of disease features relationship using Jelinik Mercer Naïve Bayes Classifier. With the objective of increasing the disease prediction rate, Jelinik Mercer Naïve Bayes Classifier is applied for disease diagnosis after minimizing the disease features.

Jelinik Mercer Naïve Bayes Classifier for RS-RMC framework is based on the Bayes rules. The Bayes rules for Classifier applies the conditional probability rule that measures the maximum likelihood of a property for the given disease features. Let us consider a scenario with a patient observed to have disease with certain symptoms (*i.e.*, features). In order to perform Multitude Classifier, Jelinik Mercer Naïve Bayes is applied to measure the relationship between the disease features and identify whether the disease being diagnosed is correct or not.

The Naïve Bayes Classifier assumes that the presence of a specific disease feature is unrelated to the presence of any other disease feature. The Naïve Bayes Classifier then efficiently predicts that given the reduct disease features “*f*”, belongs to the class “cf_i” then there exists disease feature relationship between “cf_i” and “*f*” as given as below

$$P\left(\frac{cf_i}{f}\right)=\frac{P\left(\frac{f}{cf_i}\right)*P\left(cf_i\right)}{P\left(f\right)} \tag{6}$$

where $P\left(\frac{cf_i}{f}\right)$ denotes the maximum posterior hypothesis for class “cf_i”. By applying the above formula, conditional probability of a disease pattern belonging to each disease is efficiently identified improving the class accuracy. Based on the conditional probability of disease pattern, the instance (*i.e.*, feature) is classified as the class with the highest conditional probability.

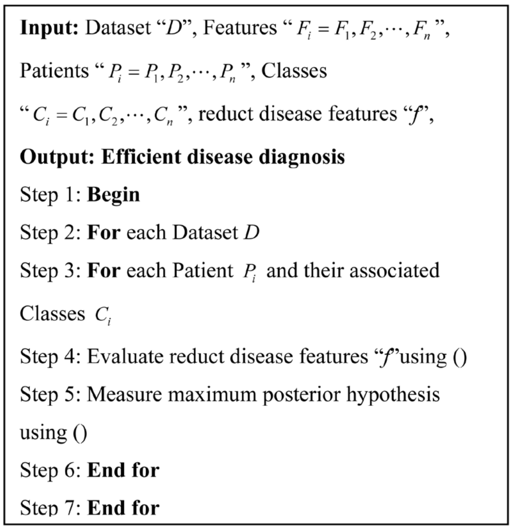
Figure 2 shows the Naïve Bayes Disease Classifier algorithm for efficient identification of classes that helps in measuring the relationship between disease features. As a result, disease diagnosis is made in an efficient manner improving the disease prediction rate. **Figure 2** shows the design of Naïve Bayes Disease Classifier algorithm. From the above algorithm, for each dataset, the features in the dataset are identified. Once the features are identified, the list of patients along with their associated classes is obtained. Based on the disease features, reduct disease features are identified to reduce the complexity without losing the values of disease feature. Finally, maximum posterior hypothesis is applied to the reduct disease features for efficient disease diagnosis.

2.3. Jelinik Mercer Multitude Classifier Model

The Jelinik Mercer in RS-RMC framework classifies Multitude Classifier for effective disease diagnosis. Jelinik Mercer Multitude Classifier is used as a smoothing technique to obtain an approximation function for multitude of disease features. It is formulated as given below

$$P\left(\frac{f}{cf_i}\right)=\left(1-\beta\right)*P\left(\frac{f}{cf_i}\right)+\beta P\left(\frac{f}{cf}\right) \tag{7}$$

From (7), “$P\left(\frac{f}{cf_i}\right)$” represent the smoothened probability of a test, given the patient medical information with existing tests and “β” ranges between “0” and “1”, “$P\left(\frac{f}{cf}\right)$” representing the maximum likelihood estimation in class feature “cf”. This in turn increases the disease prediction rate.



```
Input: Dataset "D", Features "$F_i = F_1, F_2, \cdots, F_n$",
Patients "$P_i = P_1, P_2, \cdots, P_n$", Classes
"$C_i = C_1, C_2, \cdots, C_n$", reduct disease features "f",
Output: Efficient disease diagnosis
Step 1: Begin
Step 2: For each Dataset D
Step 3: For each Patient $P_i$ and their associated
Classes $C_i$
Step 4: Evaluate reduct disease features "f"using ()
Step 5: Measure maximum posterior hypothesis
using ()
Step 6: End for
Step 7: End for
```

Figure 2. Naïve bayes disease classifier algorithm.

3. Experimental Settings

The performance of the Rough Set Rule-based Multitude Classifier (RS-RMC) framework is experimented using Cleveland Heart Disease Dataset extracted from UCI repository from Cleveland Clinic Foundation. Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/heart+Disease [10]. The data set has 76 raw attributes. However, all of the published experiments only refer to 11 of them. The RS-RMC framework is simulated using MATLAB.

The experimental work is compared against the existing Prevention and Potential Management of Ventricular Arrhythmias (PPM-VA) [1] and Prediction of Events using Spatio Spectro Temporal Data (PE-SSTD) [2] to identify the effectiveness of RS-RMC framework. The performance of the RS-RMC framework is measured in terms of disease prediction rate, execution time and false positive rate on effective disease diagnosis and class accuracy.

3.1. Execution Time

The execution time is the time taken to obtain the disease feature set. It is expressed in terms of milliseconds and is formulated as given below:

$$ET = Time(F) \tag{8}$$

From (8), "*F*" denotes the features in the Cleveland Heart Disease dataset. Lower the time taken to execute, more efficient the method is said to be

In **Figure 3**, results are reported for various classification methods for Heart Disease Dataset. On classification using PPM-VA and PE-SSTD, the execution time for obtaining two features were observed to be 0.40 ms and 0.51 ms, whereas using RS-RMC, the execution time reduced to 0.35 ms.

Figure 3 shows the time taken to obtain different features with differing sizes where features ranging from 2 to 13 were considered for experimental settings. As it can be seen the execution time steeply increases as the number of features increases, regardless of the method applied. This is because an increasing fraction of the features capacity is employed to obtain the features and diagnose disease at an early stage.

For instance with four features (*i.e.*, Chest Pain, Blood Pressure, Blood Sugar, Heart Rate), the execution time is 0.58 ms using RS-RMC framework, whereas 0.66 ms and 0.70 ms using the existing PPM-VA and PE-SSTD respectively. Moreover, the execution time is comparatively minimized using RS-RMC compared to the other methods which are demonstrated in **Figure 3**. This is because by applying Rough Set model that only selects those features that are highly essential and therefore reduces the redundant disease features. Therefore using RS-RMC the execution time for obtaining the features is reduced by 12.93% compared to PPM-VA. In a similar manner, by applying inter values to the disease features reduces the execution time by 31.45% compared to PE-SSTD.

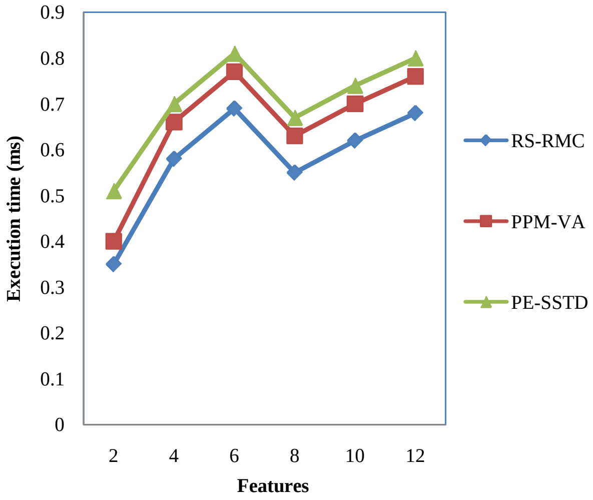


Figure 3. Measure of execution time.

3.2. False Positive Rate

The false positive rate on disease diagnosis is the ratio of absent events (*i.e.*, disease diagnosis) that yield positive test (*i.e.*, identified as disease though not) outcomes. Therefore, False Positive Rate (FPR) is the ratio of number of false positives to the total patients for conducting experiments. The mathematical formulation for FPR is given below

$$\text{FPR} = \frac{\text{No of false positives}\left(\text{identified with disease}\right)}{\text{Total patients}\left(\text{diagnosed with disease}\right)} \tag{9}$$

From (9), denotes the false positive rate. Lower the false positive rate, more efficient the method is said to be and is measured in terms of percentage (%).

Figure 4 shows the false positive rate under different simulation setting. The experiments were conducted with different number of patients and the FPR for the corresponding was measured. From **Figure 4** we can see that the value of FPR is comparatively lower in RS-RMC than the other two methods PPM-VA and PE-SSTD.

Figure 4 illustrates the impact of false positive rate and compared with two state-of-the-art works for 30 patients. **Figure 4** compares all the performance improvement based on the false positives provided by the three methods. As the number of patient increases, small rise and small fall off value is recorded irrespective of the methods used. But, RS-RMC recorded low FPR with the application of Feature Reduct in Rough Set model. By applying the Feature Reduct in Rough Set model, false positive rate for disease feature identification is reduced and therefore the disease prediction rate is also minimized. In addition by applying the Feature Reduct disease feature set is reduced without losing the information. This subsequently helps in reducing the false positive rate by 36.48% compared to PPM-VA and 74.54% compared to PE-SSTD respectively.

3.3. Disease Prediction Rate

Disease prediction rate measures the rate of disease being predicted correctly without any assumption. The Disease Prediction Rate (DPR) is the ratio of successful prediction of disease to the total number of patients and is given as below.

$$\text{DPR} = \frac{\text{Successfully predicted as disease}}{\text{Total patients}} \tag{10}$$

The performance of the different disease diagnosis method for Cleveland Heart Disease dataset is shown in **Figure 5**. It is observed that by applying RS-RMC, the disease prediction rate is increased by 6.41% to 30.64% with that of PPM-VA and PE-SSTD.

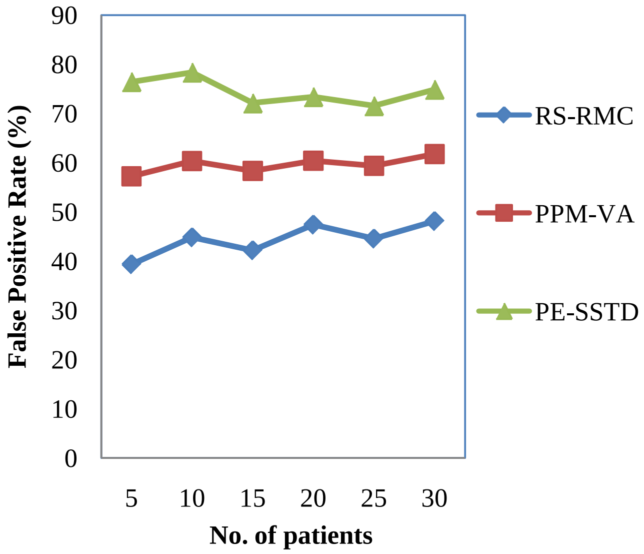


Figure 4. Measure of false positive rate with respect to patients.

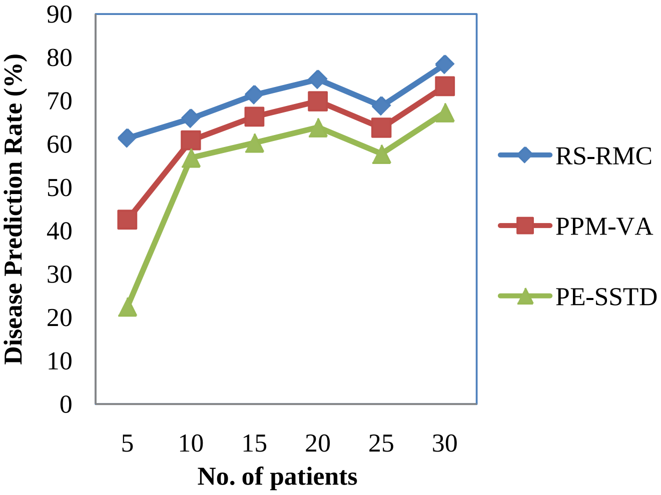


Figure 5. Measure of disease prediction rate.

Figure 5 shows the comparison of the disease prediction rate of traces with number of patients ranging from 5 to 30 and applied in Matlab. Five features were considered for obtaining the disease prediction using the three methods RS-RMC, PPM-VA and PE-SSTD respectively. From the figure, the value of disease prediction rate achieved using the proposed RS-RMC framework is higher when compared to two other existing techniques namely, PPM-VA [1] and PE-SSTD [2]. Besides we can also observe that by increasing the number of patients who provide their disease features, the disease prediction rate is increased using all the methods. But comparatively, it is higher in RS-RMC framework because the relationship between disease features is efficiently identified using Naïve Bayes Classifier algorithm. By applying Naïve Bayes Classifier algorithm, efficient identification of classes through maximum posterior hypothesis is evaluated that helps in measuring the relationship between disease features and significantly improves the disease prediction rate by 18.52% and 38.74% compared to PPM-VA and PE-SSTD respectively.

3.4. Classification Accuracy

Figure 6 shows the classification accuracy using RS-RM, PPM-VA and PE-SSTD respectively. To extract the classification accuracy, 30 patients with 20 female and 10 male patients in the age group of 40 - 55 years were considered.

Figure 6 shows the classification accuracy recorded using the three methods RS-RMC, PPM-VA and PE-SSTD. From the figure it is illustrative that the classification accuracy is improved in the proposed RS-RMC framework compared to two other methods. RS-RMC offers an improved disease diagnosis model by increasing disease prediction rate and true positive rate and decreasing the false positive rate for disease diagnosis model. Unlike the existing methods, RS-RMC used Jelinik Mercer where relationship between the features are efficiently identified and obtain approximation function for multitude of disease features. The RS-RMC framework improve its classification accuracy by reducing the execution time for obtaining the features by 55.51% and by handling over half of its disease feature set in effective disease diagnosis.

4. Conclusion

In this paper, we considered the design of a Multitude Classifier Disease Diagnosis framework to improve disease prediction rate and class accuracy in the field of medical domain is presented. A Multitude Classifier framework is introduced, and considered the problem of efficient disease diagnosis in that framework. The RS-RMC framework offers less false positive rate with lesser execution time using Rough Set Feature Selection and Feature Reduct model. Analysis of disease prediction rate demonstrates that RS-RMC framework provides higher heart disease prediction rate with the aid of Naïve Bayes Classifier algorithm. Finally, Jelinik Mercer in RS-RMC framework significantly classifies Multitude Classifier using approximation function for multitude of disease features. The performance of RS-RMC framework was compared to other disease diagnosis model,

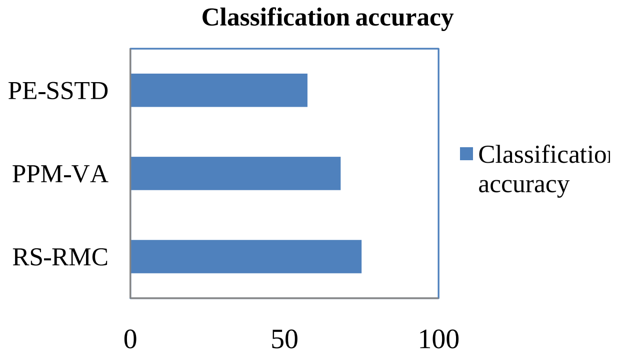


Figure 6. Measure of classification accuracy.

PPM-VA and PE-SSTD respectively. We compared the performance with many different system parameters, and evaluated the performance in terms of different metrics, such as execution time, disease prediction rate, false positive rate and classification accuracy. The results show that RS-RMC framework offers better performance with an improvement of classification accuracy by 55.51% and disease prediction rate by 28.63% compared to PPM-VA and PE-SSTD respectively.

References

[1] Koppikar, S., Baranchuk, A., Guzmán, J.C. and Morillo, C.A. (2013) Stroke and Ventricular Arrhythmias. *International Journal of Cardiology*, Elsevier, 7.

[2] Kasabov, N., Feigin, V., Hou, Z.-G., Chen, Y.X., Liang, L., Krishnamurthi, R., Othman, M. and Parmar, P. (2014) Evolving Spiking Neural Networks for Personalised Modelling, Classification and Prediction of Spatio-Temporal Patterns with a Case Study on Stroke. *Neuro Computing*, Elsevier, Vol. 134, 269-279.

[3] Meyer, K.C. (2014) Diagnosis and Management of Interstitial Lung Disease. *Springer Open Journal*.

[4] Prashanth, R., Roy, S.D., Mandal, P.K. and Ghosh, S. (2014) Automatic Classification and Prediction Models for Early Parkinson's Disease Diagnosis from SPECT Imaging. *Expert Systems with Applications*, Elsevier, **41**, 3333-3342. http://dx.doi.org/10.1016/j.eswa.2013.11.031

[5] Chawla, L.S., Herzog, C.A., Costanzo, M.R., Tumlin, J., Kellum, J.A., McCullough, P.A. and Ronco, C. (2014) Proposal for a Functional Classification System of Heart Failure in Patients with End-Stage Renal Disease. *Journal of the American College of Cardiology*, Elsevier, **63**, 1246-1252. http://dx.doi.org/10.1016/j.jacc.2014.01.020

[6] de Dios, C., Goikolea, J.M., Colomb, F., Morenoc, C. and Vietab, E. (2014) Bipolar Disorders in the New DSM-5 and ICD-11 Classifications. *Elsevier*, **7**, 179-185.

[7] Ghwanmeh, S., Mohammad, A. and Al-Ibrahim, A. (2013) Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis. *Journal of Intelligent Learning Systems and Applications*, **5**, 176-183. http://dx.doi.org/10.4236/jilsa.2013.53019

[8] Jabbar, M.A., Deekshatulu, B.L. and Chandra, P. (2013) Classification of Heart Disease Using Artificial Neural Network and Feature Subset Selection. *Global Journal of Computer Science and Technology Neural & Artificial Intelligence*, **13**, 5-14.

[9] Heckers, S., Barch, D.M., Bustillo, J., Gaebel, W., Gur, R., Malaspina, D., Owen, M.J., Schultz, S., Tandon, R., Tsuang, M., Van Os, J. and Carpenter, W. (2013) Structure of the Psychotic Disorders Classification in DSM 5. *Schizophrenia Research*, Elsevier.

[10] Frank, A. and Asuncion, A. (2010) UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine. http://archive.ics.uci.edu/ml