Binary Classification of Cognitive Disorders: Investigation on the Effects of Protein Sequence Properties in Alzheimer's and Parkinson's Disease

Shomona Gracia Jacob, Member, IAENG, Tejeswinee K

***Abstract*—Alzheimer's and Parkinson's (AD and PD) are diseases that affect the brain. They are the most common forms of dementia. It involves the degeneration of neurons in the brain cells. Diagnosis of these diseases at an early stage is said to be a challenging task and currently there exist no pre-processed database to explore the genetic role of these neuro-degenerative disorders. Hence the performance of data mining methods in extracting relevant genes that cause these neuro-degenerations is also reported to be limited. This paper targets the extraction of protein properties that are specific to AD and PD. The data is constructed and a comparative study on the performance of data mining techniques in extracting the protein properties that characterize these two ailments is investigated. The dataset is obtained by retrieving the genes related to Alzheimer's and Parkinson's disease from KEGG database. The existing data mining algorithms give a classification accuracy of around 93% with Correlation-based subset selection method. It is believed that further exploration of computational methods to investigate the role of genetic variants in causing neuro-degenerations can help identify AD and PD early. Moreover, once the genetic cause is identified, suitable drugs to target the specific gene property can well pave way towards providing cure or at least partial relief to the suffering patients.**

***Index Terms*—Alzheimer's disease, feature selection, Parkinson's disease, PROFEAT, UniProt**

I. INTRODUCTION

Data mining is the computational process of discovering potentially useful and understandable patterns in large sets involving methods of artificial intelligence, machine learning, statistics, and database systems. These patterns are useful in adding meaning to the existing data. They can also be used to predict or classify new data and in turn find the patterns inherent in them [1]. Data mining is an emerging field in providing computational diagnosis. The use of classifier systems in medical diagnosis is increasing day by day [14]. This recent advancement in technology has enabled recording of vast amounts of data.

Machine learning methods have been proposed to aid in the interpretation of such data for clinical decision making and diagnosis [2]. The existing machine learning applications fail to compete with the personalized diagnostic process done in a real clinical setting [3].

The brain is the centre of command and control in our body. Every single activity we perform is initiated by the brain. The brain composes the structural and functional properties of interconnected neurons [16]. Neurodegenerative disease is an umbrella term for a range of conditions which primarily affect the neurons in the human brain. Neurodegenerative diseases are incurable and debilitating conditions that result in progressive degeneration and/or death of nerve cells. Many of these diseases are genetic. Degenerative nerve diseases can be serious or life-threatening. It depends on the kind of ailment. Most of them have no cure. Degenerative nerve diseases affect many of our body's activities, such as balance, movement, talking, breathing, and heart function.

There are around 44 million people suffering from dementia [5]. Dementias are responsible for the greatest burden of disease with Alzheimer's representing approximately 60-70% of cases. These diseases often strike older adults. Alzheimer's disease is a degenerative disease that causes progressive/cognitive decline and memory loss. The neurons in the brain that are related to memory and language are destroyed. Alzheimer's disease affects about 1% of the people aged 65 to 69, 20% of those aged 85 to 89 years and 40% of those aged between 90 and 95 years [4]. Aging is the main factor contributing to Alzheimer's disease. Other factors may include genetics, hypertension, obesity, high cholesterol, diabetes, smoking, Down's syndrome, etc. Alzheimer's disease progresses over several years at different rates for different people. If the disease is not diagnosed at the initial stage, the severity of the disease increases. The symptoms of Alzheimer's diseases are poor decision making and judgment, misplacing things, impairments of movements, verbal communication,

Manuscript received December 7, 2016; revised December 27, 2016. This research work is a part of the Science and Engineering Research Board (SERB), Department of Science and Technology (DST) funded project under Young Scientist Scheme – Early Start-up Research Grant- titled "Investigation on the effect of Gene and Protein Mutants in the onset of Neuro-Degenerative Brain Disorders (Alzheimer's and Parkinson's disease): A Computational Study" with Reference No- SERB – YSS/2015/000737.

Dr. Shomona Gracia Jacob is a faculty in the Department of Computer Science and Engineering, SSN College of Engineering, affiliated to Anna University, Chennai, India.

Tejeswinee. K completed her B.E. in Computer Science and Engineering at Anna University, Chennai, India. Presently she is pursuing her M.E. in Computer Science and Engineering at SSN College of Engineering, affiliated to Anna University of Technology, Chennai.

abnormal moods, and complete loss of memory [6]. The diagnosis of Alzheimer's disease is done at three different stages namely, consulting the general physician, undergoing neuro-psychological tests and taking MRI scans [7]. The results of these tests confirm the onset of Alzheimer's in the subject and at this stage there is no way to reverse the damage caused. Treatments can only aim to suppress the side-effects of the disease and not cure it.

Parkinson's disease is a chronic neurological disorder based on dopamine receptors. The progressive neuro-degeneration results from the death of dopamine containing cells in the substantia nigra. It causes movement problems. It can be characterized by both motor and non-motor systems [8]. The typical features of Parkinson's disease are tremor, rigidity, bradykinesia and postural instability. The diagnosis of Parkinson's disease is usually based on the medical history and neurological investigations. The diagnosis is less accurate because of reasons such as its similarity with various other indicators and the failure to identify the disease before the subject has had a significant loss of dopamine neurons [9]. According to global declaration for PD, 6.3 million people are affected by this disease worldwide. Only 75% of clinical diagnoses of PD are confirmed to be idiopathic PD at autopsy. Thus, automatic approaches based on machine learning are needed to increase diagnosis accuracy and to help physicians make better decisions [10] and focus research on drug formulation.

A. Paper Organization

The paper is organized in the following manner: Section 2 gives a brief description of the related work. Section 3 narrates the steps involved in the process of extracting the structural and physicochemical properties along with a description of the dataset. Section 4 details the experimental results obtained using various classification algorithms while section 5 concludes the paper.

II. RELATED WORK

In the diagnosis of Alzheimer's in early step [4], the authors used medical images as an efficient tool because it provided effective assistance, both at diagnosis and therapeutic follow up. The medical images were taken from the MRI based OASIS database. They used Computer Aided Diagnostic (CAD) systems as one of the solutions to manage the medical images. For detecting the disease at an early stage, they used three sections- Hippocampus, Corpus Callosum and Cortex. Support Vector Machines (SVM) was used for the classification of data. Considering genetic factors as one of the main reasons for this disease, they stated the presence of few genes to trigger the onset of Alzheimer's. The disease progression was in three stages as: Light Stadium, Moderate Stadium and Advanced stage (terminal). The first step of the application used segmentation based on the Region of Interest (ROI) to extract the three sections. Then the SVM classifiers were applied to the extracted data. Since images were used for prediction, the diagnosis could not be early enough to take preventive measures. The disease would have progressed to a later stage where efforts can be taken to suppress the side-effects of the disease.

In the study of various machine learning techniques to detect the Alzheimer's disease at an early stage [6], the authors have compared Naive Bayes, Random Forest, JRIP and decision tree J48 algorithms. Mini-Mental State Examination (MMSE) is a popular neuro-psychological test to detect AD. But it had a disadvantage of being insensitive to the easy changes of dementia. The paper focused on diagnosis from the 10/66 battery by knowledge discovery from data. Datasets consisted of 250 patient records, divided into four age groups namely, 65-69, 70-75, 76-79 and above 80 years. Each of those groups consisted of 161, 16, 22 and 51 records respectively. There was no pre-processing step done as there was no chance of missing or incorrect values. From the dataset, 24 attributes were selected for classification. WEKA tool was used for the initial analysis of data, comparison of classification or clustering algorithms. The authors concluded that Naive Bayes classifier gave the best prediction with maximum accuracy and minimum processing time.

The paper [2] used locally weighted learning to tailor a classifier model to each patient and computed the sequence of biomarkers that were most informative or cost-effective to diagnose patients. For every patient, a personalized classifier was built by comparing the basic data about the patient with the pool data already stored. Using weights given to the attributes, with higher weights for similar subjects, a confidence level was fixed, which once attained, the diagnosis ended; otherwise the subject was studied and another biomarker was chosen. If the subject's diagnosis resulted in a confidence level higher than the described threshold, the cost of detecting was minimal. But if the process iterated for long, almost every biomarker was selected before reaching the confidence threshold. Thus the diagnosis could not be considered cost-effective in all cases.

On selecting the most influential risk factors for AD and PD using attribute evaluation scheme with ranker search method [1], the authors used chi squared feature evaluation for feature selection. The results of decision tree method were represented as classification using if-then rules. The entire work aimed at claiming; how considering the most influential risk factor for the correct classification of AD and PD can improve the accuracy of detecting the disease. The paper convinced the strong association between AD and PD. Age, genes, diabetes, smoking and stroke were the most common factors between the two diseases.

The paper talked about the application of neural networks for the detection and diagnosis of Parkinson's disease [9]. The dataset was taken from Oxford Parkinson's Disease Detection dataset. The extracted data was formatted and fed into a Multi-layered Feedforward Neural Network (MLFNN) with back propagation. The output of the MLFNN was classified with K-means clustering algorithm. Artificial Neural Network (ANN) was used. Eight attributes were chosen as the input to the ANN. All the attributes were related to voice characteristics (frequency) as tremor is a common symptom of PD. The paper concluded by

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

demonstrating an accuracy that could be increased by adding more hidden layers and choosing better weights.

A survey paper on the latest techniques for knowledge discovery using data mining techniques for identification of Parkinson's disease [11], showed diagnosis was based on the medical history and neurological examination conducted by interviewing and observing the patient in person using the UPDRS. Fisher filtering was used for feature selection. The selected features were processed to find the patterns formed by different classification algorithms. The best pattern formed was chosen based on minimum error rates. The authors verified that Random Tree classifier yielded 100% accuracy. They also suggested their method could be applied to Parkinson's Tele-monitoring dataset.

Hence, the review of previous work on AD and PD datasets has revealed the non-existence of pre-processed data for mining significant information. Moreover, this lack of proper genetic data pertaining to AD and PD has also limited the investigation of computational methods to predict the early onset of the diseases. Hence this paper presented a computational framework to explore genetic neuro-degenerative data through data mining techniques.

III. Proposed Computational Framework

A. Dataset Generation

The dataset consists of structural and physicochemical properties of proteins related to the genes of Alzheimer's and Parkinson's disease. There are 112 genes unique to both - 74 genes pertaining to Alzheimer's and 38 genes pertaining to Parkinson's disease. These genes are retrieved from the Kyoto Encyclopedia for Genes and Genomes (KEGG) database. The gene sequence for each gene was extracted from the UniProt database. One gene – LOC729317 – of Parkinson's disease did not yield an authorized gene sequence. Hence there were 111 gene instances. The structural and physicochemical properties of these genes were obtained from the PROFEAT server. The final dataset consisted of all the protein properties of 111 genes unique to Alzheimer's and Parkinson's disease.

B. Feature Selection

All the genes have 1437 protein properties each. The properties are broadly classified as G1 to G9 feature descriptors, each having a number of sub-feature categories. To find the specific genes that are most contributing to Alzheimer's and Parkinson's disease, feature selection methodologies were investigated. Three mechanisms were applied to find the optimal feature set- Correlation Feature Subset Selection (CFS), Information Gain (IG) and Gain Ratio (GR). The feature subsets obtained post feature selection were fed as input to the classification phase wherein classifiers viz, Support Vector Machine (SVM), Random Forest, Decision Tree (J48), Naive Bayes, Adaboost, k-NN were employed and their accuracy in predicting the correct diagnostic class was measured. Fig. 1 shows the proposed framework for feature selection while Fig. 2 depicts the various data mining algorithms that were investigated on the extracted feature subset.

In the framework given in Fig. 1, feature selection using Information Gain and Gain Ratio methods ranked the attributes according to their importance. The top ranked features were taken as the feature subset manually. The Correlation feature subset selection method is automated and generated a feature subset as its output. It employed the Best-First Search Strategy to identify the features that contribute the most to the prediction of the target class. Hence, there was no need for manual intervention in choosing the optimal feature set.

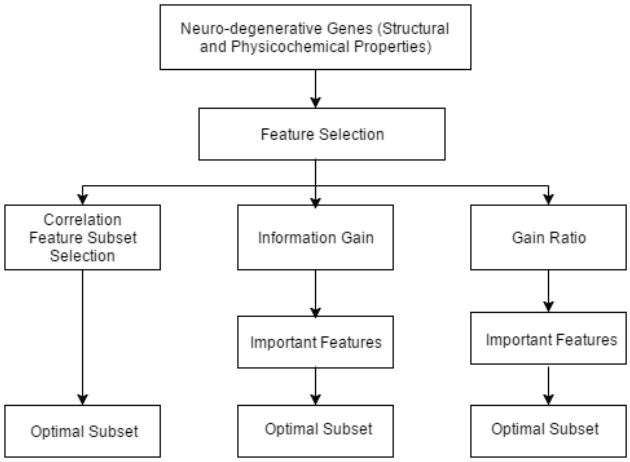

Fig. 1. Proposed Framework for Feature Selection

The classification algorithms that were explored are Support Vector Machine (SVM), Random Forest, Decision Tree (J48), Naive Bayes, Adaboost and K-NN. Their accuracy and Matthew's Correlation Coefficient (MCC) were recorded.

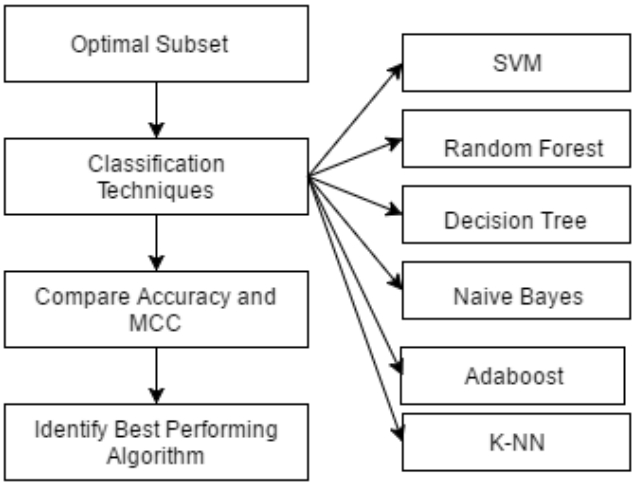

Fig. 2. Cognitive Disorder Prediction Using Extracted Feature Subset

IV. Experimental Results

The investigation of existing techniques revealed the importance of selecting important features for classification. Ten-fold cross validation was employed to measure the performance of the data mining algorithms. Two performance parameters were identified to rank the algorithms.

i. Accuracy

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard [13].

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

ii. Matthews' Correlation Co-efficient (MCC)

It's the measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives [12].

$$MCC = \frac{(TPXTN) - (FPXFN)}{\sqrt{(TP + FN)X(TN + FP)X(TP + FP)X(TN + FN)}}$$

Before feature selection (BFS), the dataset consisted of 1437 attributes and a total of 111 instances that included 74 instances of Alzheimer's disease and 37 instances of Parkinson's disease.

Investigation was carried out using WEKA [15] open source data mining suite. Once the dataset was pre-processed, feature selection techniques were explored.

A threshold greater than or equal to 0.14 was chosen for Information Gain (IG) and greater than or equal to 0.3 was chosen for Gain Ratio (GR). Correlated Feature Subset Selection (CFS) is an automated method that uses Best-First Search strategy to identify and narrow down to the optimal feature subset. The CFS subset evaluation algorithm extracted a subset containing 52 attributes. The output generated three subsets, one for each of the above mentioned mechanisms.

All the six classification algorithms were implemented and their accuracy was measured. Their results are discussed below.

TABLE I

EXPERIMENTAL RESULTS OF VARIOUS DATA MINING ALGORITHMS

CLASSIFIER	PRE-FEATURE SELECTION		POST FEATURE SELECTION					
	ACC	MCC	CFS		IG (for >=0.14)		GR (for >=0.3)	
			ACC	MCC	ACC	MCC	ACC	MCC
SVM	**0.802**	**0.544**	**0.937**	**0.857**	**0.865**	**0.689**	**0.838**	**0.629**
Random Forest	**0.820**	**0.579**	**0.892**	**0.753**	**0.847**	**0.650**	**0.856**	**0.676**
Decision Tree (J48)	0.784	0.507	0.820	0.590	0.748	0.432	0.838	0.627
Naive Bayes	0.757	0.456	0.910	0.795	0.829	0.608	0.820	0.580
Adaboost	0.766	0.445	0.847	0.645	0.811	0.559	0.820	0.582
k-NN	0.811	0.557	0.838	0.625	0.847	0.650	0.829	0.602

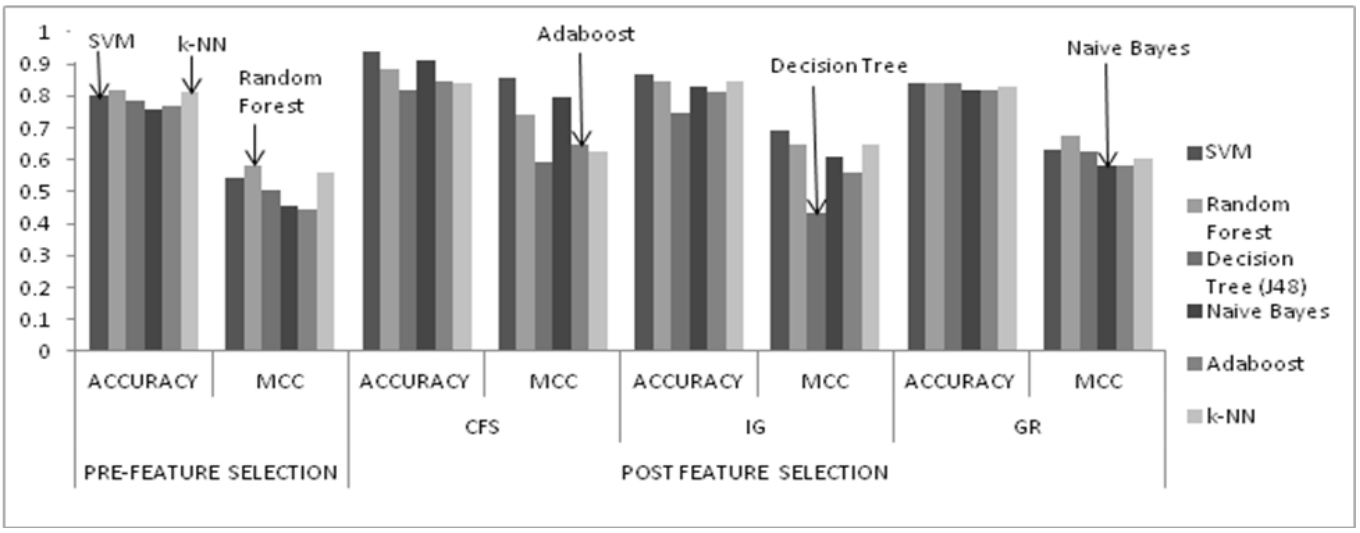

Fig. 3. Experimental Results before and after applying feature selection

Before feature selection was performed, Random forest and K-NN classification techniques give the maximum accuracy, above 80%. SVM classifier also provided considerable accuracy following the above mentioned methods.

Once the feature selection techniques were applied, from Table 1 and Fig. 3, it is proved that SVM (93.7%) and Naive Bayes (91%) classifiers outperform all the other classification algorithms with the feature subset given by CFS.

When applying Information Gain algorithm, from the ranked attributes, the attributes with a value greater than or equal to 0.14 were chosen as the subset of attributes. The subset included 40 attributes. It is also proved that SVM (86.5%) and k-NN (84.7%), Random Forest (84.7%) classifiers outperform all the other classification algorithms.

On using the Gain Ratio algorithm, from the ranked attributes, the attributes with a value greater than or equal to 0.3 were chosen as the distilled subset of attributes. The subset included 48 attributes. It is proved that Random Forest (85.6%) and SVM (83.8%) classifiers outperform all the other classification algorithms.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE II

SUMMARIZED RESULTS OF FEATURE SELECTION

FEATURE SELECTION METHOD	NO. OF FEATURES	PRE-FEATURE SELECTION		POST FEATURE SELECTION		CLASSIFIER
		MAXIMUM ACC	MCC	MAXIMUM ACC	MCC	
CFS	52	0.802	0.544	0.937	0.857	SVM
IG	40	0.802	0.544	0.865	0.689	SVM
GR	48	0.820	0.579	0.856	0.676	Random Forest

This study clearly brings out the fact that the selection of the appropriate protein properties will certainly improve the current status of prediction accuracy, in the case of neuro-degenerations.

V. CONCLUSION

Computational methods and their role in medical diagnosis have been an area of intense research in the recent past. In this paper we have derived a dataset that contains the protein properties that are prevalent in the occurrence of highly prevalent neuro-degenerations - AD and PD. The generated dataset comprised of 1437 attributes and 111 instances. The objective of this study was to investigate the performance of existing classification algorithms in extracting significant protein properties of the two diseases under study. The different classification algorithms explored in this work include SVM, Random forest, Decision tree (J48), Naive Bayes, Adaboost and K-NN. Classification accuracy was measured in terms of accuracy and Matthews Correlation Coefficient (MCC). It was evident from the study that prior to feature selection, Random forest and K-NN classifiers predicted the diagnostic classes with high accuracy (~82%) when weighed against the other classification techniques. SVM gave the best accuracy (~94%) with CFS subset evaluation. In Gain Ratio method, Random Forest showed impressive results (~85%). It was followed by SVM and Decision tree classifiers. It is evident from the investigations that selection of optimal features will certainly aid in diagnosing the disease accurately and early. Hence further research will be carried out on formulating novel methods of feature selection that would automatically select the optimal and minimal set of predictive protein properties for enhancing the diagnostic accuracy of neuro-degenerative disorders. Moreover, the identified protein properties could well be utilized for design of appropriate drugs towards curing/ providing relief to aged victims.

REFERENCES

[1] Joshi, Sandhya, et al. "Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods." *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*. IEEE, 2010.

[2] Escudero, Javier, et al. "Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease." *IEEE transactions on biomedical engineering* 60.1 (2013): 164-168.

[3] Chi, Chih-Lin, W. Nick Street, and David A. Katz. "A decision support system for cost-effective diagnosis." *Artificial Intelligence in Medicine* 50.3 (2010): 149-161.

[4] Rabeh, Amira Ben, Faouzi Benzarti, and Hamid Amiri. "Diagnosis of Alzheimer Diseases in Early Step Using SVM (Support Vector Machine)."*2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*. IEEE, 2016.

[5] Agence France Presse. (2013 December 6). "44 million now suffer from dementia worldwide". Lifestyle (Health) Magazine [Online]. Available: http://www.capitalfm.co.ke/lifestyle/2013/12/06/44-million-now-suffer-from-dementia-worldwide

[6] Shree, SR Bhagya, and H. S. Sheshadri. "An initial investigation in the diagnosis of Alzheimer's disease using various classification techniques."*Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on*. IEEE, 2014.

[7] Sheshadri, H. S. "An approach to preprocess data in the diagnosis of Alzheimer's disease." *Cloud Computing and Internet of Things (CCIOT), 2014 International Conference on*. IEEE, 2014.

[8] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." *International Journal of Bio-Science and Bio-Technology* 5.5 (2013): 241-266.

[9] Olanrewaju, Rashidah Funke, et al. "Application of neural networks in early detection and diagnosis of Parkinson's disease." *Cyber and IT Service Management (CITSM), 2014 International Conference on*. IEEE, 2014.

[10] Babu, G. Sateesh, and S. Suresh. "Parkinson's disease prediction using gene expression–A projection based learning meta-cognitive neural classifier approach." *Expert Systems with Applications* 40.5 (2013): 1519-1529.

[11] Ramani, R. Geetha, and G. Sivagami. "Parkinson disease classification using data mining algorithms." *International journal of computer applications*32.9 (2011): 17-22.

[12] Wikipedia contributors. "Matthews correlation coefficient." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 2 Dec. 2016. Web. 2 Dec. 2016.

[13] Wikipedia contributors. "Precision and recall." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Dec. 2016. Web. 3 Dec. 2016.

[14] Jacob, Shomona Gracia, and R. Geetha Ramani. "Efficient classifier for classification of prognostic breast cancer data through data mining techniques." *Proceedings of the World Congress on Engineering and Computer Science*. Vol. 1. 2012.

[15] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[16] Sahayaraj, Sanjana, and Shomona Gracia Jacob. "Knowledge Discovery through Computational Methods on EEG and fMRI Data." *International Conference on Computer Science and Application ICCSA-World Congress on Engineering and Computer Science, to be held at San Francisco*. 2015.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)