Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 115 (2017) 188–194

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India

# Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer's And Parkinson's Disease

Tejeswinee. K[a,*], Shomona Gracia Jacob[b], Athilakshmi. R[c]

[a]*Rajalakshmi Engineering College, Thandalam, Chennai – 602105, India*
[b]*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai – 603110, India*
[c]*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai – 603110, India*

## Abstract

Alzheimer's and Parkinson's disease are the most common forms of dementia that degenerate neurons in the brain cells. This paper targets a comparative study on the performance of data mining techniques in neuro-degenerative data. The existing data mining algorithms give classification accuracy ~93% with Correlation-based feature subset selection method. The proposed Decremental Feature Selection Method has yielded a more optimal feature subset that gives higher accuracy in prediction. Further exploration of computational methods to investigate the role of such genetic variants will aid in identifying the genetic cause of these diseases and design suitable drugs to target the gene property.

Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications.

*Keywords:* Alzheimer's; Parkinson's; Feature Selection; UniProt; PROFEAT

## 1. Introduction

Alzheimer's disease and Parkinson's disease are the diseases that affect the brain. Alzheimer's disease is a degenerative disease that causes progressive decline of cognition and memory. It causes the degeneration of the nerve cells in the brain related to memory and language. Symptoms appear after 65 years and the prevalence increases sharply with age. It is the most common form of dementia [1]. Genetic factors play an important role in the

* Corresponding author. Tel.:+91-9791088650.
*E-mail address:* tejeswinee1530@cse.ssn.edu.in

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications
10.1016/j.procs.2017.09.125

onset of the disease as certain genes can increase the suffering, although they are directly not the cause of the disease. Other factors include age, smoking, and alcohol intake [2]. The symptoms of Alzheimer's diseases are poor decision making and judgment, misplacing things, impairments of movements, verbal communication, abnormal moods, and complete loss of memory[13],[14],[15]. If the disease is not diagnosed at the initial stage, the severity of the disease increases. The diagnosis of AD is done at three different stages namely, consulting the General Physician, undergoing neuropsychological tests and taking MRI scans [3]. The current diagnostic tools like MMSE (Mini-Mental State Examination) and NFT (Neurofibrillary Tangles) scores have poor sensitivity especially for the early stages of the disease and do not allow for diagnosis until the disease has lead to irreversible brain damage. Diagnosing the disease at its earlier stage can be difficult due to the similarity in clinical symptoms of rare degenerative dementias [4].

Parkinson's disease is a neurological disorder based on dopamine receptors. It affects the mobility of the subject. It is a progressive condition characterized by both motor and non-motor symptoms. People with Parkinson's disease are presented with the symptoms and signs associated with Parkinsonism, namely, hypokinesia, rigidity, bradykinesia and rest tremor. It is also caused by drugs and less common conditions such as multiple cerebral infraction and degenerative conditions such as Progressive Supra nuclear Palsy (PSP) and Multiple System Atrophy (MSA) [5]. Diagnosis depends on the presence of two or more cardinal motor features. Having so many factors to analyze to diagnose the disease, the specialist normally makes decisions by evaluating the current test results of their patients. Also, the previous decisions made on other patients with a similar condition are reviewed [6]. Such procedures are complex, especially when the number of factors that the specialist has to evaluate is high.

Data mining technology is being adopted in biomedical sciences and research for providing prognosis and deep understanding of the classification of disease [7]. The use of classifier systems in medical diagnosis is increasing day by day. This recent advancement in technology has enabled recording of vast amounts of data [8]. Machine learning methods have been proposed to aid in the interpretation of such data for clinical decision making and diagnosis. Most of these methods achieved promising prediction accuracies; however they were evaluated on different pathologically unproven data sets making it difficult to make a fair comparison among them. Also, many other factors such as pre-processing, the number of important attributes for feature selection, class imbalance distinctively affect the assessment of the prediction accuracy [17]. A model proposed with an initial pre-processing step followed by imperative attribute selection and classification can achieve satisfying results [16], [18].

## 2. Proposed Computational Framework

This section describes the computational framework adopted for the investigation on neuro-degenerative disorders. The authors have collected this data and it is to be stated that this data is being reported for the first time and has not been earlier utilized for computational explorations.

### *2.1. Dataset Generation*

Dataset generation begins with collecting the genes related to Alzheimer's and Parkinson's disease from Kyoto Encyclopedia of Genes and Genomes [19] (KEGG) database. A total of 112 genes are collected. There are 74 genes uniquely pertaining to Alzheimer's disease and 38 genes uniquely pertaining to Parkinson's disease. There are 95 genes common to both the diseases.

The gene sequences for every gene were obtained in the next step from the UniProt database. A gene related to Parkinson's disease – LOC729317 – had no authorized gene sequence and hence was disregarded and not included in further investigations. The output at this phase gives the gene sequence for 111 genes.

The next phase deals with the extraction of structural and physicochemical properties of the 111 genes from the PROFEAT server. There are 1437 features or protein properties for every gene. Thus the dataset consists of 111 rows of genes with 1437 columns representing the protein properties. The class label that identifies the disease takes the toll to1438 columns in the final dataset.

### 2.2. Feature Selection

Utilizing all the 1437 features for disease classification is a time-intensive task. Further, a few genes may contribute more towards either Alzheimer's or Parkinson's disease while the rest may not do as much. Hence, identification of the genes which have the potential to highly influence the triggering of either or both of the diseases helps in recording higher classification accuracy.

The 1437 protein properties are grouped into 9 classes labeled as feature descriptors from G1to G9. Each of the 9 feature descriptors has a number of sub-feature categories [20].

Three feature selection methodologies were studied for the selection of optimal feature set. Correlation Feature Subset Selection (CFS), Information Gain (IG) and Gain Ratio (GR) were deployed individually to the complete dataset. The result obtained was the optimized feature subset. This subset was individually fed to each of the six classification algorithms in the classification phase. The algorithms investigated were Support Vector Machine (SVM), Random Forest, Decision Tree, Naïve Bayes, Adaboost and Nearest Neighbor (K-NN). The accuracy of all these algorithms in predicting the correct diagnostic class was measured. Figure 1 shows the proposed framework for feature selection.

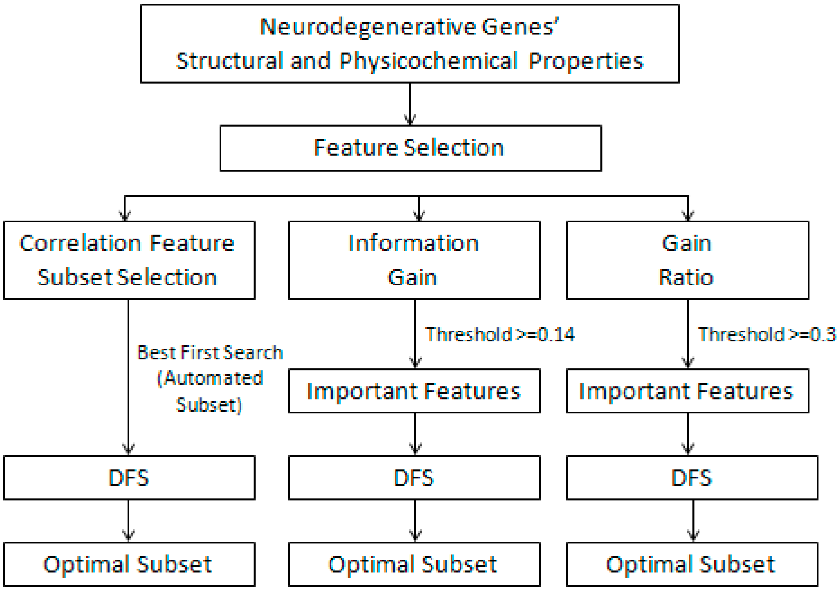

Fig.1. Proposed Framework for Feature Selection

In Figure 1, feature selection method using Information Gain and Gain Ratio methods ranked the attributes according to their importance. The top ranked features were taken as the optimized subset manually. The threshold for selection was based on the size of the returning feature subset. Correlation Feature Subset selection method is automated and produced an optimized feature subset. It used best-first search strategy to identify the features that contribute the most towards the prediction of the target class.

## 3. Experimental Results

This section briefs on the investigations carried out and the results obtained.

Exploring the six data mining algorithms validated the importance of feature selection. Ten-fold cross validation was employed to measure the performance of data mining algorithms. Two performance parameters were identified to rank the algorithms.

i. Accuracy

The degree to which the result of a measurement, calculation or specification conforms to the correct value or action [9].

ii. Matthew's Correlation Coefficient

It is the measure of quality of binary (two-class) classifications. It takes into account true and false positives and negatives [10].

Prior to performing feature selection, the dataset consisted of 1437 features and 111 instances in which 74 instances voted for Alzheimer's disease and 37 for Parkinson's disease. Investigations were carried out with WEKA open source data mining suite [11]. The dataset was pre-processed and the data mining algorithms were implemented on the pre-processed dataset. To each of the six algorithms, the input from the three feature selection methods viz. Correlation Feature Subset selection, Information Gain and Gain Ratio were applied and the results were tabulated.

Correlation Feature Subset selection method being automated, selected a subset consisting of 52 features. A threshold value greater than or equal to 0.14 was chosen for Information Gain and a value greater than or equal to 0.3 was chosen for Gain Ratio method. The threshold was decided based on the number of features that fell within the range (~50) [21]. Table 1 discusses the results of all the techniques.

Table 1. Experimental Results of Various Data Mining Algorithms

| Classifier | Pre-Feature Selection | | Post Feature Selection | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | CFS | | IG (for >=0.14) | | GR (for >=0.3) | |
| | ACC | MCC | 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 | 0.784 | 0.507 | 0.820 | 0.590 | 0.748 | 0.432 | 0.838 | 0.637 |
| 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 |

Prior to feature selection, Random Forest and K-NN methods gave a maximum accuracy of above 80%. SVM classifier provided considerable accuracy following the above methods. Post feature selection, as evident from Table 1 and Figure 2, it is proved that SVM (93.7%) and Naïve 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 features with a value greater than or equal to 0.14 were chosen as the subset attributes. The subset included 40 attributes. It was observed that SVM (86.5%), K-NN (84.7%) and Random Forest (84.7%) yielded the highest accuracy with the Information gain subset.

On using Gain Ratio algorithm, from the ranked attributes, the features with a value greater than or equal to 0.3 were chosen as the distilled subset of attributes. The subset included 48 attributes. It was noted that Random Forest (85.6%) and SVM (83.8%) classifiers outperformed all the other classification algorithms [21]. Table 2 shows the summarized results of feature selection.

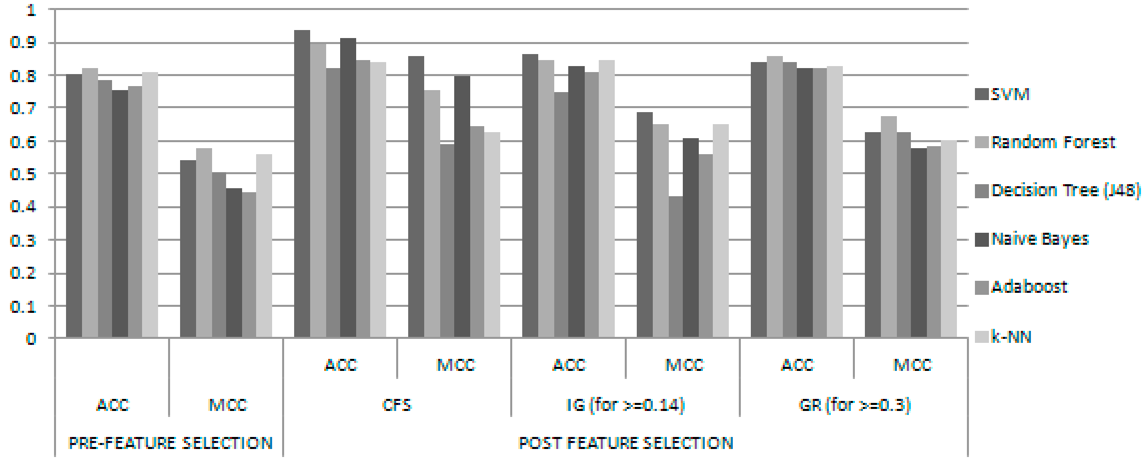

Fig. 2. Experimental Results Before and After Applying Feature Selection

Table 2. Summarized Results of Feature Selection

| Feature Selection Method | No. Of Features | Pre-Feature Selection | | Post Feature Selection | | Classifier |
|---|---|---|---|---|---|---|
| | | Max. ACC | MCC | Max. ACC | MCC | |
| CFS | 54 | 0.802 | 0.544 | 0.937 | 0.857 | SVM |
| IG | 42 | 0.802 | 0.544 | 0.865 | 0.689 | SVM |
| GR | 50 | 0.820 | 0.579 | 0.856 | 0.676 | Random Forest |

Ranking algorithms presented the features in the order of their contribution to categorizing the samples under the two target classes [12]. Decremental Feature Selection (DFS) method was applied in order to investigate if a more optimal feature subset could be identified from the results of the three methods – IG, GR, CFS. This method further reduced the number of attributes required for the accurate prediction of the class. An attribute per iteration was removed, starting from the last attribute in the feature subset. The best performing classification algorithm was again executed on the new subset of features and the accuracy was noted. It was observed that the accuracy increased steadily as features were removed before becoming stable. The subset of attributes which recorded the maximum accuracy was recognized as the final set of optimized attributes. Table 3 shows the results after DFS was executed.

Table 3. Summarized Results after Decremental Feature Selection

| Feature Selection Method | No. of Features before DFS | Max. ACC | MCC | No. of Features after DFS | Max. ACC | MCC | Classifier |
|---|---|---|---|---|---|---|---|
| CFS | 54 | 0.937 | 0.857 | 49 | **0.973** | **0.940** | **SVM** |
| IG | 42 | 0.865 | 0.689 | 37 | 0.874 | 0.713 | SVM |
| GR | 50 | 0.856 | 0.676 | 36 | 0.883 | 0.743 | Random Forest |

From Table 3, it is proved that the Accuracy and MCC increased when DFS was applied over each of the three feature selection methods. When SVM classifier was implemented, the accuracy increased from 93.7% to 97.3%

under CFS feature selection. Information Gain technique when applied along with SVM classifier showed a raise in accuracy from 86.5% to 87.4%. Similarly, Random Forest classifier employed over Gain Ratio method increased the accuracy from 85.6% to 88.3%. Also, the number of features in the feature subset decreased from 54, 42 and 50 to 49, 37 and 36 for CFS, IG and GR techniques respectively.

## 4. Conclusion

The application of data mining in the field of medical research has been an area of tremendous interest. In the recent past, using genetic data for disease diagnosis has gained momentum. It has propelled the study and identification of diseases at their early stage. In this paper, we have generated a new dataset that consists of genetic information that pertains to the two neuro-degenerative disorders – Alzheimer's and Parkinson's disease. The dataset, initially extracted from the KEGG database, consisted of 1437 structural and physicochemical protein properties that were extracted from the PROFEAT server. It consisted of 111 instances where 74 genes that uniquely contributed to Alzheimer's and 37 genes that distinctly contributed to Parkinson's disease. The objective of this investigation was to examine the performance of the classification algorithms such as Support Vector Machine (SVM), Random Forest, Decision Tree, Naïve Bayes, Adaboost and K-NN, on the generated dataset. The accuracy of classification by these algorithms was measured using two units – Accuracy and Matthew's Correlation Coefficient (MCC) – and the results were tabulated. 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 noticeable from the study that selecting the optimal features by CFS followed by implementing DFS, thus creating a hybrid feature selection technique, improves the classification accuracy and aids in early disease diagnosis. These results can be utilized for designing appropriate drugs towards providing relief to affected subjects.

## Acknowledgements

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.

## References

[1] A.B. Rabeh, *et al.* Diagnosis of Alzheimer Diseases in Early Step Using SVM (Support Vector Machine). *Computer Graphics, Imaging and Visualization (CGiV)* 2016: 364-367.

[2] S.R. Bhagyashree, & H.S. Sheshadri, An approach in the diagnosis of Alzheimer disease—a survey. *Int J Eng Trends Technol (IJETT)* 2014; ***7*(1)**: 41-43.

[3] Michael saling, *et al.* Early Diagnosis of Dementia. *Alzheimer's Australia* 2007.

[4] S. Joseph, *et al.* A Statistical and Biological Approach for identifying misdiagnosis of incipient Alzheimer patients Using Gene expression Data. In *Engineering in Medicine and Biology Society* 2006: 5854-5857.

[5] S. Bind, *et al*. A survey of machine learning based approaches for parkinson disease prediction. *International Journal of Computer Science and Information Technologies* 2015; ***6*(2):**1648-1655.

[6] R. Ramani, *et al.* Parkinson disease classification using data mining algorithms. *International journal of computer applications* 2011: 17-22.

[7] S. Joshi, *et al*. Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods. In *Machine Learning and Computing (ICMLC)* 2010: 218-222.

[8] J. Escudero, *et al*. Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease. *IEEE transactions on biomedical engineering* 2013; **60(1)**: 164-168.

[9] D. Gupta, K. Berberich. Diversifying search results using time 2016.

[10] S. Harding, *et al*. Cartesian genetic programming for image processing. In *Genetic programming theory and practice X* 2013: 31-44.

[11] M. Hall, *et al.* The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 2013; **11(1)**: 10-18.

[12] R.G. Ramani, S.G. Jacob. Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models. *PloS one* 2013; **8(3)**: e58772.

[13] S.G. Jacob, R. Athilakshmi. Extraction of Protein Sequence features for Prediction of Neuro-degenerative Brain Disorders: Pioneering the

CGAP database. In *Proceedings of the International Conference on Informatics and Analytics* 2016: 30.

[14] S.B. Shree, H.S. Sheshadri. Diagnosis of Alzheimer’s Disease using Rule based Approach. *Indian Journal of Science and Technology* 2016; **9(13)**.

[15] S.B. Shree, H.S. Sheshadri. An initial investigation in the diagnosis of Alzheimer's disease using various classification techniques. In *Computational Intelligence and Computing Research (ICCIC)* 2014: 1-5.

[16] A. Khan, M. Usman. Early diagnosis of Alzheimer's disease using machine learning techniques: A review paper. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)* 2015: 380-387.

[17] R. Chaves, *et al.* Effective diagnosis of Alzheimer’s disease by means of association rules. *Hybrid Artificial Intelligence Systems* 2010: 452-459.

[18] A. Khan, *et al*. Early Diagnosis of Alzheimer Disease Using Instance-Based Learning Techniques. *Journal of Medical Imaging and Health Informatics* 2016; **6(4)**: 1111-1118.

[19] M. Kanehisa. The KEGG database. *silico simulation of biological processes* 2002; **247**: 91-103.

[20] E. Boutet, *et al*. UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Plant bioinformatics: methods and protocols* 2007: 89-112.

[21] K. Tejeswinee, S.G. Jacob. Binary classification of Cognitive Disorders: Investigation on the Effects of Protein Sequence Properties in Alzheimer's and Parkinson's Disease. *IAENG-IMECS* 2017: 166-170.