Classification of Alzheimer's Disease using Machine Learning Techniques

Muhammad Shahbaz[1], Shahzad Ali[2][a], Aziz Guergachi[3], Aneeta Niazi[1] and Amina Umer[1]
[1]*Department of Computer Science and Engineering, University of Engineering and Technology, Lahore-54890, Pakistan*
[2]*Department of Information Technology, University of Education Lahore, Multan Campus, Multan, Pakistan*
[3]*Department of Information Technology Management, TRS, Ryerson University, Toronto, ON, Canada*

Keywords: Alzheimer's Disease, Data Mining, Machine Learning, Healthcare, Classification, Neurodegeneration.

Abstract: Alzheimer's disease (AD) is a commonly known and widespread neurodegenerative disease which causes cognitive impairment. Although in medicine and healthcare areas, it is one of the frequently studied diseases of the nervous system despite that it has no cure or any way to slow or stop its progression. However, there are different options (drug or non-drug options) that may help to treat symptoms of the AD at its different stages to improve the patient's quality of life. As the AD progresses with time, the patients at its different stages need to be treated differently. For that purpose, the early detection and classification of the stages of the AD can be very helpful for the treatment of symptoms of the disease. On the other hand, the use of computing resources in healthcare departments is continuously increasing and it is becoming the norm to record the patient' data electronically that was traditionally recorded on paper-based forms. This yield increased access to a large number of electronic health records (EHRs). Machine learning, and data mining techniques can be applied to these EHRs to enhance the quality and productivity of medicine and healthcare centers. In this paper, six different machine learning and data mining algorithms including k-nearest neighbours (k-NN), decision tree (DT), rule induction, Naive Bayes, generalized linear model (GLM) and deep learning algorithm are applied on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset in order to classify the five different stages of the AD and to identify the most distinguishing attribute for each stage of the AD among ADNI dataset. The results of the study revealed that the GLM can efficiently classify the stages of the AD with an accuracy of 88.24% on the test dataset. The results also revealed these techniques can be successfully used in medicine and healthcare for the early detection and diagnosis of the disease.

1 INTRODUCTION

Neurodegenerative, continuous deterioration of neurons, diseases are usually considered as a group of disorders that damage the working competence of the human nervous system intensely and progressively (Scatena et al., 2007). Alzheimer's disease (AD) is the key public health concern throughout the world and one of the most widespread neurodegenerative disorder (Small, 2005). The AD is a cureless disease because it has no diagnosis and treatment methods to slow its progression or stop its onset (Unay et al., 2010). The median survival duration of the patients suffering from AD has been estimated to be only 3.1 years for the initial diagnosis of the probable AD, and only 3.5 years for the initial diagnosis of possible AD (Wolfson et al., 2001).

The frequency of AD occurrence is becoming more common in late life, especially the people at the age of greater than 65 are at high risk (Cummings et al., 2014). At present, there are approximately 44 million victims of AD dementia throughout the world. If the breakthroughs fail to identify the prevention and diagnosis of the AD, it is expected to be increased to a number of more than 100 million by the year 2050 (Association et al., 2013; Touhy et al., 2014). During the AD, degeneration of the brain progresses with time. Therefore, the patients suffering from the AD should be categorized into different subgroups, depending upon the stage of the

[a] https://orcid.org/0000-0002-0608-9515

Shahbaz, M., Ali, S., Guergachi, A., Niazi, A. and Umer, A.
Classification of Alzheimer's Disease using Machine Learning Techniques.
DOI: 10.5220/0007949902960303
In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*, pages 296-303
ISBN: 978-989-758-377-3

disease. This division is critical because the patients at different stages of the AD are required to be treated differently, and the same medication cannot be used for all of them (Gamberger et al., 2017). For that purpose, the classification of different stages of the AD can be very helpful for the treatment of symptoms of the disease to improve the patient's quality of life.

The use of computing resources in healthcare departments is continuously increasing and it is becoming the norm to record the patient data electronically that was traditionally recorded on paper-based forms. This yield increased access to many electronic health records (EHRs) but 80% of the data is unstructured. As a result, the processing of the unstructured data is challenging and difficult using database management software and other traditional methods. Machine learning and data mining tools and techniques can be applied to these EHRs to enhance the quality and productivity of medicine and healthcare centers (Alonso et al., 2018). Data mining or knowledge discovery is the practice of finding the unknown and useful patterns from many pre-existing datasets. Such patterns are used to understand the historical dataset, to classify new data and generate summaries of data. In this way, data mining helps in the discovery of the deeper patterns in data, as well as classify or group records on the basis of similarities or dissimilarities between them (Sumathi and Sivanandam, 2006).

Since the last few decades, data mining has been extensively used in many areas such as marketing, retail, banking, stock market prediction, and medicine and healthcare, etc. (Agarwal et al., 2018; Canlas, 2009; Thenmozhi and Deepika, 2014; Yang et al., 2018). The algorithms used for data mining can be applied to group different subjects based on the similarities in their attributes (Eapen, 2004). Many studies (Soni and Gandhi, 2010; Dipnall et al., 2016; Ni et al., 2014 have shown that machine learning, and data mining techniques have been significantly used in medicine and healthcare research. The application of data mining in medicine helps in the efficient prognosis of several diseases, understanding the classification of disease, specifically in neuroscience, and biomedicine. The main aim of this research work is to apply machine learning and data mining techniques on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset to classify the different stages of the AD. Moreover, the sub-objective of this research work is to identify the most distinguishing attribute for each of the different stages of the AD among ADNI dataset.

2 DATA

2.1 Data Description

In this research work, the dataset has been collected from The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) challenge which is available at https://tadpole.grand-challenge.org. It is the dataset recorded from the North American individuals, who participated in ADNI. ADNI is the multicentre study aimed to improve the clinical, genetic, biochemical, and imaging biomarkers for early diagnosis of the AD. A standard set of procedures and protocols is followed during ADNI data collection, to avoid any inconsistencies in the data (Weiner et al., 2013). The TADPOLE data has been recorded for both male and female participants, including mild cognitive impairment subjects, old aged individuals, and AD patients. The data contain records of participants' examination carried out at 62 different sites, at different monthly intervals, ranging from baseline (i.e. 0 months) to 120 months, from July 2005 to May 2017.

2.2 Data Exploration

The original TADPOLE dataset contains 1,907 attributes for 1,737 participants. On the basis of diagnosis, these participants have been divided into five different classes namely Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), Subjective Memory Complaint (SMC), and Alzheimer's Disease (AD). The number of participants and their examination records for each of the five stages of AD included in the TADPOLE dataset are illustrated in Table 1.

Table 1: Original TADPOLE dataset details.

Class	No. of Participants	No. of Records
CN	417	3,821
EMCI	310	2,319
LMCI	562	4,644
SMC	106	389
AD	342	1,568

TADPOLE dataset contains 12,741 examination records of 1,737 participants. The dataset is unevenly distributed among the five classes of participants, i.e. LMCI and CN have more data as compared to AD, EMCI, and SMC. The class SMC has the least data, with just 106 participants having 389 examination records (Table 1).

2.3 Data Pre-Processing and Feature Selection

In this research work, a subset of TADPOLE dataset has been used. The examination records of 530 participants (106 participants from each AD stage) have been selected out of 1,737 participants, to perform experiments using different data mining algorithms for the classification of five different stages of the AD. In data mining techniques, the completeness of data is very critical for precise modeling. While the TADPOLE dataset contains a lot of sparseness. Therefore, we have selected only those attributes, which have maximum data coverage. An attribute will have complete data coverage if it will have a value for all the 12,741 instances of examination records. But, there are only 41 attributes, out of 1,907 attributes, which have a value for more than 10,000 instances of examination records.

Furthermore, an analysis has been carried out to remove redundancy from the 41 attributes having enough data. There are 6 attributes in the TADPOLE dataset, that indicate the time of participant's examination, including EXAMDATE (date of examination), EXAMDATE_bl (date of first examination), M (months of examination, to nearest 6 months as continuous), Month (months of examination, to nearest 6 months as a factor), Month_bl (fractional value of months of examination), and VISCODE (participant's visit code, indicating months of examination). As all these attributes indicate the time of the participant's examination. In this way, only one attribute namely VISCODE has been selected for indicating the participant's examination date, and all other five redundant attributes have been removed. Similarly, the dataset contains two attributes for the participant's identification, including RID (participant Roaster ID) and PTID (Participant ID). The attribute PTID has been removed from the dataset to avoid redundant IDs. After data analysis and pre-processing, 28 attributes have been selected for this investigation.

These attributes are divided into three categories. The first category is demographics attributes which include the general quantifiable characteristics of participants. The second category is cognitive assessment attributes which include the attributes representing the cognitive behavior of a participant. For that purpose, different cognitive assessment tests are carried out and scores are assigned to the participant based on their cognitive abilities. Finally, the third category is clinical assessment attributes which include the significant biomarkers of the AD. These attributes have been recorded after the clinical examination of all participants. The list of demographic, cognitive assessment and clinical assessment attributes with their description that have been extracted for analysis is illustrated in Table 2, 3 and 4, respectively.

Table 2: List of demographic attributes in the dataset used in this analysis.

Label	Description	Data Type	Units
RID	Participant's Roaster ID.	Numeric	NA
AGE	Participant's age.	Numeric	Years
PTGENDER	Participant's gender	Nominal	NA
PTEDUCAT	Participant's education	Numeric	Years
PTETHCAT	Participant's ethnicity	Nominal	NA
PTRACCAT	Participant's race	Nominal	NA
PTMARRY	Participant's marital status.	Nominal	NA
VISCODE	Participant's Visit code.	Nominal	NA
Years_bl	Participant's year of examination.	Numeric	Years
SITE	A code indicating the site of a participant's examination	Numeric	NA

Table 3: List of cognitive assessment attributes in the dataset used in this analysis.

Label	Description	Data Type	Units
CDRSB_bl	Clinical Dementia Rating Sum of Boxes (core)	Numeric	NA
ADAS11_bl	11 item-AD Cognitive Scale (score)	Numeric	NA
ADAS13_bl	13 item-AD Cognitive Scale (score)	Numeric	NA
MMSE_bl	Mini-Mental State Examination (score)	Numeric	NA
RAVLT_immediate_bl, RAVLT_learning_bl, RAVLT_forgetting_bl, RAVLT_perc_forgetting_bl	Rey's Auditory Verbal Learning Test (scores for immediate response, learning, forgetting and percentage forgetting)	Numeric	NA
FAQ_bl	Functional Activities Questionnaire	Numeric	NA

Table 4: List of clinical assessment attributes in the dataset used in this analysis.

Label	Description	Data Type	Units
APOE4	APOE4 gene presence	Binary	NA
Hippocampus_bl	Volume of hippocampus	Numeric	mm^3
Ventricles_bl	Volume of ventricles	Numeric	mm^3
WholeBrain_bl	volume of Brain	Numeric	mm^3
Fusiform_bl	The volume of the fusiform gyrus.	Numeric	mm^3
Entorhinal_bl	The volume of the entorhinal cortex.	Numeric	mm^3
MidTemp_bl	The volume of the middle temporal gyrus.	Numeric	mm^3
ICV	Intra Cranial Volume	Numeric	mm^3

2.4 Data Partitioning

The dataset is divided into a training dataset (70% of the entire dataset) and test datasets (30% of the entire dataset). According to this division, the 2,164 examination records are used for training of the data mining models, and the 927 examination records are used for model testing. The data counts of training and test dataset after pre-processing are given in Table 5. It illustrates that all the five classes have a different number of examination records. Because in TADPOLE dataset, the number of examination records is not the same for all participants.

Table 5: Data Counts of each AD stage for the Training and Test Dataset.

AD Stage	Training Samples	Test Samples
CN	632	270
EMCI	448	192
LMCI	566	243
SMC	217	93
AD	301	129

3 METHODOLOGY

In this research work, six different machine learning and data mining algorithms including K-nearest neighbors (K-NN), decision tree (DT), rule induction, Naive Bayes, generalized linear model (GLM) and deep learning algorithm are applied on ADNI dataset in order to classify the five different stages of the AD. In this investigation, rapidminer studio, one of the famous data mining tools, is used for implementing all these algorithms.

3.1 K-Nearest Neighbours (K-NN)

K-nearest neighbor algorithm is a simple data mining technique used for both classification and regression problems. K-NN classification algorithm assigns an object to a specific class based on the majority classes of its K neighbors. The value of K, a positive integer, defines the number of neighbors to be considered for polling (Zhang and Zhou, 2005). In this analysis, the value of K is set to 11 which is selected using the trial and error method.

3.2 Decision Tree (DT)

The decision tree algorithm is a predictive modeling technique commonly used for classification in data mining, statistics, and machine learning applications. It classifies the dataset by computing the information gain values for all attributes of a dataset. The leaf nodes in a tree denote a class label while the branches to these leaf nodes denote the combination of input variables that lead to those class labels (Shahbaz et al., 2013).

3.3 Rule Induction

The rule induction is a data mining algorithm, in which a pruned set of rules is extracted from the training data, based on maximum values of information gain. The rules are in the form of 'if-then' statements (see Appendix B). Rule sets have an advantage over decision trees, as they are simpler to comprehend and can be represented in first-order logic (Stepanova et al., 2018).

3.4 Naive Bayes Algorithm

The Naive Bayes algorithm is one of the data mining and machine learning classification technique which is based on the Bayesian theorem. It used to find the probability of an attribute based on other known probabilities that are related to the attribute. It uses Gaussian probability densities for data modeling (Thomas and Princy, 2016).

3.5 Generalized Linear Model (GLM)

The generalized linear model (GLM) is a supervised machine learning approach used for both classification and regression problems. It is an extension of traditional linear models. GLM classifies the data based on the maximum likelihood between attributes. It performs parallel computations and is an extremely fast machine learning approach and works very well for models with a limited number of predictors (Guisan et al., 2002).

3.6 Deep Learning

Deep learning models are one of the machine learning techniques. It is based on the multilayer feedforward artificial neural networks, which are vaguely inspired by the working of the biological nervous system or the human brain. In this investigation, a deep neural network with two hidden layers is trained using backpropagation learning algorithm. The detailed working of deep neural networks is given in (Levine et al., 2018). The schematic representation of the proposed methodology is illustrated in Figure 1.

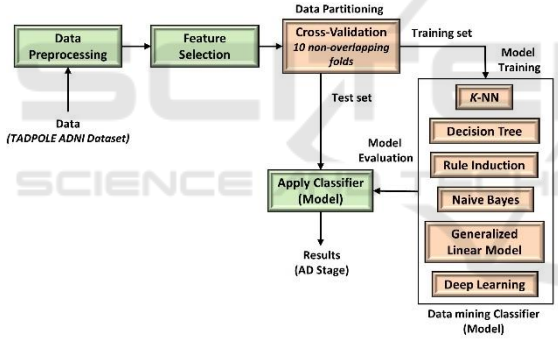

Figure 1: Schematic representation of the proposed methodology.

3.7 Model Evaluation – Confusion Matrix

The performance of the classifier is usually evaluated using a confusion matrix. It is a specific table which is used to presents the true classes and classifier predicted classes and illustrates the type of errors made by the classifier. Confusion matrix for binary classification (class '0' and class '1') is shown in Fig. 2. There are four different terminologies used in the confusion matrix which are given as follows:

- **True Positive** (TP) means classifier predicted positive ('1') and it is true ('1').
- **True Negative** (TN) means classifier predicted negative ('0') and it is true ('0').
- **False Positive** (FP) means classifier predicted positive ('1') and it is false ('0').
- **False Negative** (FN) means classifier predicted negative ('0') and it is false ('1').

	True '1'	True '0'
Predicted '1'	TP	FP
Predicted '0'	FN	TN

Figure 2: Confusion matrix for binary classification.

Accuracy, precision, and recall are the important performance metrics that are calculated from a confusion matrix for a classification model.

Accuracy means how often the classification model correctly classifies the data samples?

$$Accuracy = \frac{TP+TN}{Total} \quad (1)$$

Precision is the number of TP classes over the sum of both the TP classes and FP classes.

$$Precisio = \frac{TP}{TP+ FP} \quad (2)$$

Recall is the number of TP classes over the sum of both the TP classes and FN classes.

$$Recall = \frac{TP}{TP+ FN} \quad (3)$$

4 RESULTS AND DISCUSSIONS

In this investigation, all the classification models are trained with the 10 folds cross-validation of models on the training dataset. Cross-validation is performed with the training dataset, to avoid overfitting of the models. The performance of the classifiers is evaluated using the unseen test dataset. The classification accuracy of the classifiers during the validation period and the test period is illustrated in Table 6.

Table 6: The classification accuracy of data mining classifiers during Validation and test period.

Classifier	Accuracy (%)	
	Validation	Test
K-NN	73.10	43.26
Decision Tree	76.43	74.22
Rule Induction	92.47	69.69
Naive Bayes	79.44	74.65
Generalized Linear Model	92.75	88.24
Deep Learning	78.79	78.32

Table 7: Confusion matrix obtained by applying a generalized linear model to the test dataset.

	True CN	True LMCI	True EMCI	True SMC	True AD	Precision
Pred. CN	255	6	0	25	0	89.16%
Pred. LMCI	0	209	4	0	0	98.12%
Pred. EMCI	0	28	174	17	0	79.45%
Pred. SMC	15	0	14	51	0	63.75%
Pred. AD	0	0	0	0	129	100 %
Recall	94.44%	86.01%	90.62%	54.84%	100 %	

The results analysis indicates that the generalized linear model outperforms other classifiers and gives an accuracy of 88.24% during the test period. Reasonable accuracies are obtained for deep learning and Naive Bayes algorithms i.e. 78.32% and 74.65% respectively. It can also be observed that the results obtained on test data are quite close to the results obtained from cross-validation period. This shows that the developed models are not overfitted during their training period. From the decision tree and rule induction models, it has been observed that the most distinguishing attribute for the five stages of the AD is the CDRSB cognitive test, as it appears at the top of the decision tree. The most distinguishing clinical assessment attribute is the volume of the whole brain, whereas the most distinguishing demographic attribute is the age of the patient (see Appendices A and B). The detailed results obtained by applying the generalized linear algorithm during the test period is illustrated in Table 7.

It can be observed that generalized linear model correctly classified most of the unseen instances of AD, CN, EMCI and LMCI classes, with the class recall of 100.00%, 94.44%, 90.62%, and 86.01% respectively, and class precision of 100.00%, 89.16%, 98.12%, and 79.45% respectively. The results of the SMC class show misclassification of testing instances. It can be observed from the above table that 25 out of 93 instances of SMC class have been misclassified with the CN class. This is because the patients belonging to SMC class have very similar values of clinical assessment attributes as those who belong to CN class. As well as, 28 out of 243 instances of the LMCI class have been misclassified as EMCI. This is because the attribute values of EMCI and LMCI classes overlap with each other.

5 CONCLUSION AND FUTURE DIRECTIONS

Machine learning, and data mining techniques are very helpful in medicine and healthcare studies for early detection and diagnosis of several diseases. The results of the research work illustrate that accuracy of the GLM is 88.24% for the test period. The results also indicate that the most distinguishing attributes for the different stages of AD include the CDRSB cognitive test among the cognitive assessment attributes, the volume of the whole brain among the clinical assessment attributes and age of the patient among the demographic attributes. Furthermore, the results proved that the machine learning, and data mining techniques can be successfully used in the early detection, prediction, and diagnosis of several diseases. The accuracy of the AD stages classification could be further improved by increasing the number of instances for EMCI and SMC classes so that the model can be trained with sufficient and balanced data for all classes.

ACKNOWLEDGMENTS

The authors are thankful to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data committee for providing access to their AD dataset.

REFERENCES

Agarwal, N., Koti, S. R., Saran, S., and Kumar, A. S. (2018). Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for new delhi, india. CURRENT SCIENCE, 114(11):2281–2291.

Alonso, S. G., De La Torre-D´ıez, I., Hamrioui, S., L´opez-Coronado, M., Barreno, D. C., Nozaleda, L. M., and Franco, M. (2018). Data mining algorithms and techniques in mental health: a systematic review. Journal of medical systems, 42(9):161.

Association, A. et al. (2013). 2013 alzheimer's disease facts and figures. Alzheimer's & dementia, 9(2):208–245.

Canlas, R. (2009). Data mining in healthcare: Current applications and issues. School of Information Systems & Management, Carnegie Mellon University, Australia.

Cummings, J. L., Morstorf, T., and Zhong, K. (2014). Alzheimer's disease drug-development pipeline: few

candidates, frequent failures. Alzheimer's research & therapy, 6(4):37.

Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., and Meyer, D. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. PloS one, 11(2): e0148195.

Eapen, A. G. (2004). Application of data mining in medical applications. Master's thesis, University of Waterloo.

Gamberger, D., Lavra˘c, N., Srivatsa, S., Tanzi, R. E., and Doraiswamy, P. M. (2017). Identification of clusters of rapid and slow decliners among subjects at risk for alzheimer's disease. Scientific reports, 7(1):6763.

Guisan, A., Edwards Jr, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological modelling, 157(2-3):89–100.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International Journal of Robotics Research, 37(4-5):421–436.

Ni, H., Yang, X., Fang, C., Guo, Y., Xu, M., and He, Y. (2014). Data mining-based study on sub-mentally healthy state among residents in eight provinces and cities in china. Journal of Traditional Chinese Medicine, 34(4):511–517.

Scatena, R., Martorana, G. E., Bottoni, P., Botta, G., Pastore, P., and Giardina, B. (2007). An update on pharmacological approaches to neurodegenerative diseases. Expert opinion on investigational drugs, 16(1):59–72.

Shahbaz, M., Guergachi, A., Noreen, A., and Shaheen, M. (2013). A data mining approach to recognize objects in satellite images to predict natural resources. In IAENG Transactions on Engineering Technologies, pages 215–230. Springer.

Small, D. H. (2005). Acetylcholinesterase inhibitors for the treatment of dementia in alzheimer's disease: do we need new inhibitors? Expert opinion on emerging drugs, 10(4):817–825.

Soni, N. and Gandhi, C. (2010). Application of data mining to health care. International Journal of Computer Science and its Applications.

Stepanova, D., Gad-Elrab, M. H., and Ho, V. T. (2018). Rule induction and reasoning over knowledge graphs. In Reasoning Web International Summer School, pages 142–172. Springer.

Sumathi, S. and Sivanandam, S. (2006). Introduction to data mining principles. Introduction to data mining and its applications, pages 1–20.

Thenmozhi, K. and Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. International Journal of Engineering Research and General Science, 2(6):6–11.

Thomas, J. and Princy, R. T. (2016). Human heart disease prediction system using data mining techniques. In 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pages 1–5. IEEE.

Touhy, T. A., Jett, K. F., Boscart, V., and McCleary, L. (2014). Ebersole and Hess' Gerontological Nursing and Healthy Aging, Canadian Edition-E-Book. Elsevier Health Sciences.

Unay, D., Ekin, A., and Jasinschi, R. S. (2010). Local structure-based region-of-interest retrieval in brain mr images. IEEE Transactions on Information technology in Biomedicine, 14(4):897–903.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al. (2013). The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimer's & Dementia, 9(5):e111–e194.

Wolfson, C., Wolfson, D. B., Asgharian, M., M'lan, C. E., Østbye, T., Rockwood, K., and Hogan, D. f. (2001). A reevaluation of the duration of survival after the onset of dementia. New England Journal of Medicine, 344(15):1111–1116.

Yang, J., Li, J., and Xu, Q. (2018). A highly efficient big data mining algorithm based on stock market. International Journal of Grid and High Performance Computing (IJGHPC), 10(2):14–33.

Zhang, M.-L. and Zhou, Z.-H. (2005). A k-nearest neighbour based algorithm for multi-label classification. In Granular Computing, 2005 IEEE International Conference on, volume 2, pages 718–721. IEEE.

APPENDIX

Decision Tree Model

CDRSB_bl>0.250
| CDRSB_bl>2.750
| | MMSE_bl>26.500
| | | AGE>75.300: LMCI {AD=0, CN=0, EMCI=0, LMCI=13, SMC=0}
| | | AGE≤75.300: EMCI {AD=0, CN=0, EMCI=8, LMCI=0, SMC=0}
| | MMSE_bl≤26.500: AD {AD=254, CN=0, EMCI=5, LMCI=0, SMC=0}
| CDRSB_bl≤2.750
| | MMSE_bl>23.500
| | | RAVLT_perc_forgetting_bl>-9.167
| | | | AGE>88.850: AD {AD=5, CN=0, EMCI=0, LMCI=0, SMC=0}
| | | | AGE≤88.850
| | | | | Fusiform_bl>23133.500
| | | | | | PTETHCAT = Hisp/Latino: EMCI {AD=0, CN=0, EMCI=2, LMCI=0, SMC=0}
| | | | | | PTETHCAT = Not Hisp/Latino: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=5}
| | | | | Fusiform_bl≤23133.500
| | | | | | Ventricles_bl>9770.500
| | | | | | | ICV_bl>1268370

| | | | | | | | RAVLT_immediate_bl>52.500: EMCI {AD=0, CN=0, EMCI=78, LMCI=0, SMC=5}
| | | | | | | | RAVLT_immediate_bl≤52.500: LMCI {AD=23, CN=32, EMCI=344, LMCI=553, SMC=3}
| | | | | | | | SITE≤64.500: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=4}
| | | | | | Ventricles_bl≤9770.500
| | | | | | | SITE>67.500: EMCI {AD=0, CN=0, EMCI=6, LMCI=0, SMC=0}
| | | | | | | SITE≤67.500: CN {AD=0, CN=14, EMCI=0, LMCI=0, SMC=0}
| | | RAVLT_perc_forgetting_bl≤-9.167: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=5}
| | MMSE_bl≤23.500: AD {AD=19, CN=0, EMCI=0, LMCI=0, SMC=0}
CDRSB_bl≤0.250
| AGE>69.550
| | SITE>147: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=6}
| | SITE≤147
| | | WholeBrain_bl>1234525: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=6}
| | | WholeBrain_bl≤1234525
| | | | FAQ_bl>2.500: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=5}
| | | | FAQ_bl≤2.500
| | | | | ADAS13_bl>19.500: SMC {AD=0, CN=0, CN=1, EMCI=0, LMCI=0, SMC=5}
| | | | | | | | ICV_bl≤1814440: CN {AD=0, CN=566, EMCI=0, LMCI=0, SMC=77}
| | | | | | | RAVLT_learning_bl≤0.500: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=3}
| | | | | | MMSE_bl≤25: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=4}
| AGE≤69.550
| | PTEDUCAT>10.500
| | | AGE>64.050: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=83}
| | | AGE≤64.050
| | | | PTGENDER = Female: SMC {AD=0, CN=0, EMCI=0, LMCI=0, SMC=2}
| | | | PTGENDER = Male: CN {AD=0, CN=12, EMCI=0, LMCI=0, SMC=0}
| | PTEDUCAT≤10.500: CN {AD=0, CN=7, EMCI=0, LMCI=0, SMC=0}

Rule Induction Model

if CDRSB_bl≤0.750 and Fusiform_bl≤16443.500 and ADAS13_bl≤15.500 and RAVLT_immediate_bl>32 then CN (0/282/7/0/16)
if RAVLT_perc_forgetting_bl>86.607 and MMSE_bl>26.500 and ADAS13_bl>19.500 then LMCI (0/0/4/131/0)
if MidTemp_bl>20054 and CDRSB_bl>0.250 and ICV_bl≤1697385 and RAVLT_forgetting_bl>3.500 then EMCI (0/0/169/4/1)
if Years_bl>2.471 and FAQ_bl>0.500 and SITE≤55 then LMCI (6/0/6/121/0)
if CDRSB_bl≤0.750 and ICV_bl>1489935 and Years_bl>0.990 and RAVLT_immediate_bl>41.500 then CN (0/137/0/0/14)
if Hippocampus_bl≤7242.500 and FAQ_bl≤4.500 and Entorhinal_bl≤3037 and Fusiform_bl≤16003.500 then LMCI (0/0/0/66/0)
if MMSE_bl≤26.500 and FAQ_bl>4.500 then AD (285 / 0 / 0 / 25 / 0)
fif CDRSB_bl≤0.750 and ICV_bl>1562475 and WholeBrain_bl≤1121605 and PTGENDER = Male then CN (0/106/0/1/0)
if CDRSB_bl>0.750 and ADAS11_bl≤8.500 then LMCI (0/0/3/42/0)
if ADAS13_bl≤13.500 and SITE≤34.500 and MidTemp_bl>20055.500 then SMC (0/5/0/0/48)
if ADAS13_bl>13.500 and RAVLT_learning_bl>4.500 and FAQ_bl >0.500 then EMCI (0/0/40/0/0)
if ADAS11_bl>12.500 then LMCI (0/3/0/37/0)
if Ventricles_bl≤28863.500 and MMSE_bl>28.500 and AGE>70.600 then CN (0/56/0/0/3)
if Fusiform_bl>18434 and Fusiform_bl≤18888.500 then EMCI (0/0/36/0/2)
if FAQ_bl>1.500 and MidTemp_bl≤19842 then LMCI (0/0/0/36/0)
if MidTemp_bl≤19923.500 and WholeBrain_bl>977838 then SMC (0/1/0/0/41)
if Ventricles_bl≤8152 then CN (0/19/0/0/2)
if PTEDUCAT≤14.500 and ADAS13_bl>17.500 then EMCI (0/0/24/0/0)
if Ventricles_bl≤30608 and CDRSB_bl>0.250 then LMCI (0/0/0/32/0)
if AGE≤70.250 then SMC (0/1/0/0/17)
if ADAS13_bl≤11.500 and Ventricles_bl>40723 then CN (0/17/0/0/0)
if Hippocampus_bl>8041 then EMCI (0/0/13/0/0)
if ADAS13_bl>17.500 then AD (10/0/0/0/0)
if AGE≤78 then SMC (0/0/0/0/8)
else LMCI (0/0/0/1/0)