Predicting the Likelihood of Alzheimer’s Disease from Clinical and Magnetic Resonance Image Data

Ji Yoon Song

**Predicting the Likelihood of Alzheimer’s Disease from Clinical & Magnetic Resonance Image Data**

Ji Yoon Song

**ABSTRACT**

Alzheimer’s disease (AD) is a form of dementia found more commonly in aging populations. Although the exact cause and mechanism of the disease are subjects of an ongoing investigation, it is possible to use key demographic and clinical data in conjunction with features extracted from MRI scans to predict with reasonable accuracy if a subject is likely to develop Alzheimer’s disease.

Public data made available by OASIS (Open Access Series of Imaging Studies) was used, which consists of cross-sectional collection of 416 subjects aged 18 to 96, and longitudinal collection of 150 subjects aged 60 to 96. Data provided include, per subject, demographic data such as gender, age, education level, socioeconomic status, and clinical data, such as MMSE (Mini-Mental State Examination, Rubin et al., 1998), and CDR (Clinical Dementia Rating, Morris, 1993), and also derived anatomic volumes data: eTIV (Estimated Total Intracranial Volume), ASF (Atlas Scaling Factor), nWBV (Normalized Whole Brain Volume).

Due to the complex nature of the data, several machine learning classifiers were trained and their accuracies measured. The provided data was divided into training/test sets, and the training set was processed with machine learning classifiers including Logistic Regression, MLP (Multilayer Perceptrons), KNN, Gaussian Naïve Bayes, Random Forest, and SVM. The test set was then applied to the classifiers, and the accuracies of the individual filters were measured using k-fold Cross Validation. With existing data, accuracy levels attained were on average 58% to 76%, with a maximum of around 82%. This demonstrates that building a prediction model useful in detection of Alzheimer’s Disease is possible. With increased number of subjects to train these filters, and other key features, it may be possible to increase the accuracy even further.

1. **INTRODUCTION**

Alzheimer’s disease (AD), the most common form of dementia, is associated with progressive mental deterioration that results in cognitive impairments, and is known to be incurable and irreversible. It is also a global phenomenon: in 2006 26.6 million people were affected by AD, and in 2050 this is projected to quadruple, resulting in 1 person in 85 living with the disease1. This potentially life-threatening disease is caused by abnormal accumulation of Amyloid beta, forming amyloid plaques and intracellular neurofibrillary tangles, with widespread gliosis, loss of synapses and degeneration of neurons2. A definitive diagnosis can be made post-mortem tests for the presence of these elements in the patient’s brain, but with protracted diseases as AD it is essential that the patient is discovered and treated early on, reducing occurrences of potential life-threatening incidents.

Many researchers have been working on an effective and accurate way of detecting AD early on, among which structural imaging based on magnetic resonance has played a key part. Magnetic Resonance Image (MRI) scans can reveal atrophy of medial temporal structures, which is considered a valid diagnostic marker at the Mild Cognitive Impairment (MCI) stage3. Specifically, researches have concluded that even mild AD is associated with atrophy of cortical association areas, and that the degree of clinical impairment is associated with the severity of observed regional atrophy4. With MRI scans, these regions of atrophy can be visually recognized by trained medical personnel, but for exploratory data analysis a more numerical approach is needed. Efforts have been made to quantize this degree of atrophy by calculating brain volumes from MRI scans. One major difficulty was that of correcting for head size variation in brain morphometric analyses, and a novel method to resolve this was to adjust total intracranial volume (TIV) with the Atlas Scaling Factor (ASF) to scale each individual and arrive at normalized values5. The resulting metric estimated total intracranial volume (eTIV) can be used as an indicator for brain atrophy along with ASF and normalized whole brain volume (nWBV).

To ensure clinical examinations can discover AD early on, there are more conventional methods that help diagnose AD. The Mini-Mental State Examination (MMSE) is one such tool, that can be used to systematically and thoroughly assess mental status. This is a 11-question measure that tests five areas of cognitive function: orientation, registration, attention and calculation, recall, and language. The maximum score is 30, and a score of 23 or lower indicates cognitive impairment6. Although it has been mentioned that there is no evidence supporting MMSE as a stand-alone test in the identification of MCI patients who could develop dementia including AD, with repeated tests and other diagnostic tools this could help early identification of potential patients7. Another indicator that may have bearing is the socioeconomic status (SES), and educational level. Lower education level does suggest higher risk of AD, but this is not directly mediated by SES. However, low SES at an earlier age was associated with increased risk8. It is also worth noting that not all studies have revealed association between education level and dementia. A more consistent relationship with dementia occurs when years of education reflects cognitive capacity, which suggests that effect of education on risk of dementia should be evaluated within the context of a lifespan developmental model9. It is also popular belief that women have more AD than men, and that significantly more people develop AD in older age than young, which suggests that these parameters should be considered when building a predictive model on AD.

Since definitive diagnosis of AD cannot be performed until the brain actually has the symptoms of the disease, by which time it may be too late to help the patient, it seems to be logical to focus detection on the early stages of AD or MCI. The Clinical Dementia Rating (CDR), developed by the Memory and Aging Project at Washington University School of Medicine, is a 5-point scale used to characterize six domains of cognitive and functional performance applicable to AD10. A score of 0 indicates that the subject is normal, while 0.5 indicates very mild dementia and 1 indicates mild dementia. This may provide very useful information to predict AD dementia progression in MCI individuals11. In the analysis presented in this paper, CDR is used as the target value for the predictive model.

The complex nature of potential indicators of AD, ranging from quantitative MRI measures, clinical and demographic tools, suggests that using machine learning is efficient. Machine learning techniques have successfully been applied in studies linking biomarkers to obtain relevant relationships that explain different stages in AD12. In this paper, we use a variety of machine learning classifiers that can learn to predict whether an individual is suffering from clinical dementia, test them for accuracy using statistical analysis, and look for insights on which variable contributes most to successful prediction.

1. **MATERIAL AND METHODS**
   1. Data Source

The Open Access Series of Imaging Studies (OASIS) project offers MRI datasets of the brain for free distribution. The data used in this paper was downloaded directly from the OASIS website (<http://www.oasis-brains.org/>). This public data consists of two sets:

* + - * Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults: “This set consists of a cross-sectional collection of 416 subjects aged 18 to 96. … The subjects are all right-handed and include both men and women. 100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer’s disease (AD) …”
      * Longitudinal MRI Data in Nondemented and Demented Older Adults: “This set consists of a longitudinal collection of 150 subjects aged 60 to 96. … The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits …”

Each set includes, for each individual, multiple raw scan images, averaged image, atlas-registered gain field-corrected image and also a brain-masked version of that image.

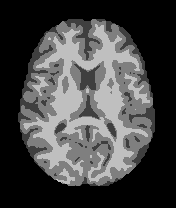


Figure 1. Sample grey/white/CSF segmentation image generated from the masked atlas image

Since this paper does not aim to focus solely on image data for building predictive models, it was more efficient to resort to derived data presented for each dataset. The data includes, in the form of a CSV file13:

* + - * “**Demographics**: Gender (M/F), Handedness (Hand), Age, Education (Educ), socioeconomic status (SES) (Rubin et al.,1998). Education codes correspond to the following levels of education: 1: less than high school grad., 2: high school grad., 3: some college, 4: college grad., 5: beyond college.”
      * “**Clinical**: Mini-Mental State Examination (MMSE) (Rubin et al., 1998), Clinical Dementia Rating (CDR; 0=nondemented; 0.5=very mild dementia; 1=mild dementia; 2=moderate dementia) (Morris, 1993). All participants with dementia (CDR >0) were diagnosed with probable AD.”
      * “**Derived anatomic volumes**: Estimated total intracranial volume (eTIV) (mm3) (Buckner et al., 2004), Atlas scaling factor (ASF) (Buckner et al., 2004), Normalized whole brain volume (nWBV) (Fotenos et al., 2004).”
  1. Data Processing & Load

**Platform**: The analysis was done using Python 3.6, using NumPy, Pandas and Scikit-learn libraries ([Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011). The platform was a Mac running an installation of Anaconda.

**Data Paring**: The original data on both sets contained data that had no demographic/clinical columns, and were discarded. This was necessary because without CDR rating the predictive model will assume that the subject is nondemented (CDR=0), and will result in confusion because the subject may or may not in reality be demented. Any subject partially missing columns were left blank. The original files provided by OASIS were “oasis\_longitudinal.csv” and “oasis\_cross-sectional.csv”. The modified files were renamed “oasis\_longitudinal2.csv” and “oasis\_cross-sectional2.csv”. After paring the data analysis was done on the remaining 233 cross-sectional cases and 370 longitudinal cases.

**Data Preparation Process**: The raw CSV files were read into Pandas dataframes. Each column in the dataframe was then was processed using LabelEncoder in ‘sklearn.preprocessing’ which encodes categorical integer features using a one-hot aka one-of-K scheme. This is used to convert string labels into integers, but will also map real numbers into a comparable range of integers. This approach works better with most classifiers.

* 1. Machine Learning Classifiers

**Logistic Regression**: Logistic Regression is an approach to learning functions of the form f: X →Y, or P(Y|X) in the case where Y is discrete-valued, and X is any vector containing discrete or continuous variables. P(X) follows the form of the logistic function. Logistic Regression is a closely related alternative to Gaussian Naïve Bayes (GNB), though the two can produce different results in many cases14. In the Scikit-learn implementation, the multiclass case uses one-vs-rest (OVR) scheme.

**Multilayer Perceptrons (MLP)**: MLP uses multiple layers of Perceptrons feedforwarding from input to hidden to output layers. A part of classical neural network theory, uses stochastic gradient descent and backpropagation to allow vectors of weights to transform input data into a space where it becomes linearly separable15. The existence of hidden layers allows MLPs to become a universal approximator. The downside to MLP is that it can be costly in terms of system resources compared to other more modern methods.

**K-Nearest Neighbor (KNN)**: This is a type of instance-based learning or non-generalizing learning. It doesn’t generate an internal model but instead stores instances of the training data. Classification is calculated from majority vote of the nearest neighbors of each point (<http://scikit-learn.org>). A small k value is more flexible, with low bias but high variance. A larger k value is resistant to outliers with lower variance.

**Gaussian Naïve Bayes (GNB)**: Naïve Bayes reduces the intractable sample complexity inherent in Bayesian classifiers by making a conditional independence assumption that dramatically reduces the number of parameters to be estimated while modeling P(X|Y), from exponential to linear16. Despite simplified assumptions, GNB works quite well in real world situations, requiring only a small amount of training data to estimate necessary parameters. GNB can be extremely fast compared to more sophisticated methods, and works well even if the parameters are dependent (H. Zhang (2004). [The optimality of Naive Bayes.](http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf) Proc. FLAIRS).

**Random Forest Decision Tree**: The random forest algorithm is a perturb-and-combine technique designed for trees. This means a diverse set of classifiers is created by introducing randomness in the classifier construction. The generalization error for forests converges as the number of trees in the forest becomes large. It yields error rates that compare favorably to Adaboost (Freund and Schapire[1996]) but are more robust towards noise (L. Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001).

**Support Vector Classifier (SVC)**: SVC is the Scikit-learn implementation of Support Vector Machines (SVM), using the library libsvm (<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>). A non-probabilistic classifier, SVM is practically one of the most effective classification algorithms in machine learning. Key concepts are margins and kernels. Margin is the distance from a data point to the linear separator, and SVM directly optimizes for the maximum margin separator. If the data is not linearly separable, as in the case of pixel data, a kernel ‘trick’ is used, implicitly mapping inputs into high-dimensional feature spaces17. The Scikit-learn implementation of SVC uses the ‘one-vs-one’ approach for multi-class classification (Knerr et al., 1990). If n is the number of classes, then n\*(n-1)/2 classifiers are constructed and each one trains data from two classes (<http://scikit-learn.org/stable/modules/svm.html#svm-classification>).

1. **RESULTS**

Applying the preprocessed data to each of the previously discussed machine learning filters, it was necessary to use particular measures of accuracy, to overcome the problem of overfitting. Overfitting occurs when a model would have a perfect score on the samples that it has seen, but would fail to predict anything on the data it has not seen. Typically learning the parameters of a prediction function and testing it on the same data would generate this error. To avoid overfitting, it is customary to split the available data into two sets, a test set and a training set. For the model to generalize to an independent dataset, cross-validation is used.

Two methods were chosen to test each machine learning classifier:

**Repeated Random Sub-Sampling Validation (aka Monte Carlo)**: This approach uses random splits, but reiterates the split so that each time a different pair of training and testing subsets are created. The advantages of this method is that the split is not dependent on the number of iterations. In this study 30 iterations were used for each run, and the five stat summary was drawn into a boxplot. Implementation: The sklearn.model\_selection method train\_test\_split() function was used to create a random split of the data used into training set(70%) and test set(30%). Then matplotlib was used to show the result of 30 iterations as a boxplot.

**K-fold Cross Validation**: This approach partitions the original data into k equal sized random subsamples. Then one subsample is held as test data, and the remaining k-1 subsamples are used to train the classifier. This cross-validation process is repeated k times (folds). Each of the subsamples are used for testing exactly once. Then the k results are averaged to produce a single accuracy score. Implementation: sklearn.model\_selection has an implemented version of this method. The function cross\_val\_score() was used to create a list of k scores for each fold. Then the list was subjected to a boxplot. The difficult part was determining optimal value for k. This was different for each classifier, and an optimal number had to be found by comparing the average score for each value of k. For some classifiers, this number seemed to converge (i.e. not change over iterations), but for others each iteration of cross validation changed the optimal value. In this case a conservative value of 10 was used.



Table 1. Summary of results for longitudinal dataset. Legends (MC:MonteCarlo, CV:k-fold CrossValidaton, mean:mean, st.d:standard deviation)

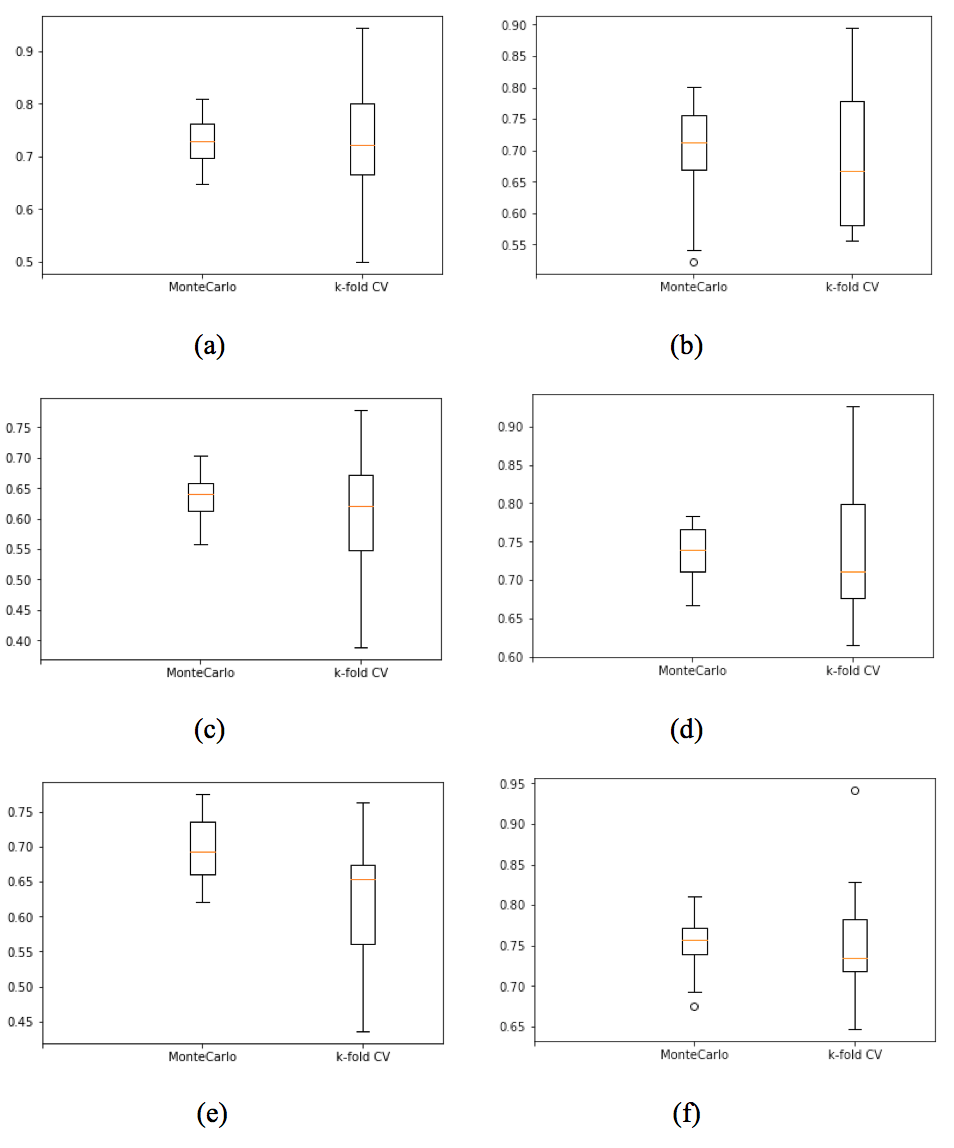


Figure 2. Boxplots for Table 1 (longitudinal dataset). Legends (a: Logistic Regression, b: MLP, c: KNN (2), d: GBN, e: Random Forest, f: SVC)



Table 2. Summary of results for cross-sectional dataset. Legends (MC:MonteCarlo, CV: k-fold Cross Validaton, mean: mean, st.d: standard deviation)

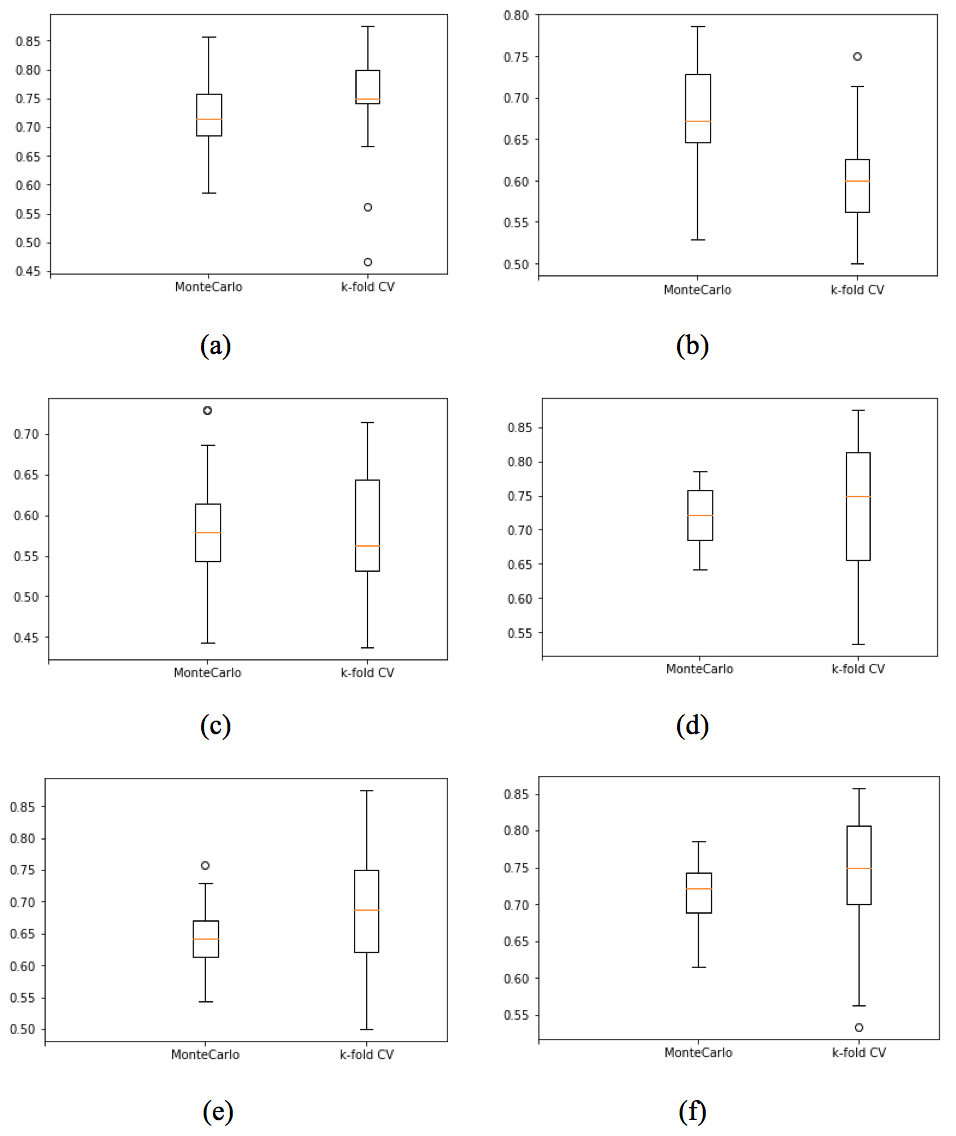


Figure 3. Boxplots for Table 2 (cross-sectional dataset). Legends (a: Logistic Regression, b: MLP, c: KNN (2), d: GBN, e: Random Forest, f: SVC)

1. **DISCUSSION**

In determining the best classifier to use for this experiment, the Monte Carlo method was relatively unstable, the average accuracy followed a distribution centered somewhere near the value attained by k-fold CV method. While this is to be expected, since the selection process is randomized, this makes it difficult to come up with a single figure that represents performance. So The numbers from k-fold CV method, almost constant across iterations, were used to determine the classifier with the highest accuracy. Overall SVC is the most accurate; in the longitudinal data it is the clear winner with 76%, while in the cross-sectional dataset it is a very close second (73%), after a tied first position by Logistic Regression and GNB (74%). The actual difference is less than 1% due to round-off error. The worst classifier overall was KNN (2) on cross-sectional dataset (58%). Increasing the number of neighbors only reduced the accuracy; the current choice yielded the best performance overall. KNN was also the lowest performer on the longitudinal dataset.

After cleansing the data of subjects with missing feature values, the longitudinal dataset shows approximately similar relative performances to the cross-sectional dataset. Overall, the accuracy is a bit lower for the cross-sectional dataset, possibly owing to the fact that it has a greater variability for a key feature, Age. In theory, this should improve the performance of kernel based SVMs, for example, but a closer look at the data shows that almost all of the additional subjects with younger age were nondemented. This has the effect of clustering the subset with older subjects relatively closer together, making it harder for SVM to converge to a state with higher accuracy. Also, after the data paring was complete, there were less subjects used in the experiment for the cross-sectional dataset.

A closer look at the features was taken after the main results were complete. In an easy to replicate mini-experiment, the features that were used were edited. Each of the features were taken singly, to reduce the model to a one-feature problem. Surprisingly, MMSE alone created an SVC model with 74% accuracy, only 2% lower than with all other features combined. This may be due to the fact that both MMSE and CDR are derived from tests that are designed to detect cognitive impairment. It also means that the neuromorphic numbers (eTIV, nWBV, ASF) by themselves do not contribute as much as expected. This probably is due to the fact that the output variable is set to CDR, which indicates in our dataset’s range up to mild dementia. At that point the brain’s actual structural atrophy should not have progressed much, leading to smaller deviations in the neuromorphic features. This is within the confines of this paper’s assumptions, because the purpose is to determine possibility of AD before the disease has progressed far.

On the analysis of the accuracy data, since the experiment used CDR as the y variable, which has in the pared dataset three possible values (0, 0.5, and 1\*), it was a multiclass classification task, not a binary classification. Therefore, binary performance indicators such as sensitivity and specificity were not included in the analysis. The overall accuracy seems to be lower than optimal, but considering the dataset poses a three class classification [[1]](#footnote-1)task the result is rather significant. In another experiment using the same dataset and k-SVM classifier, even with preprocessing that included pixel information extracted from the actual image the accuracy rose to 74%~80%18. To increase the accuracy further, either or both of these two measures should be taken: 1) turn this into a binary classification problem, by creating an additional column indicating dementia. CDR value of 0 should be classified nondemented, and anything higher should be classified demented. This should have the added benefit of being able to include the three subjects with CDR rating of 2. 2) analyze and extract key features from the actual image, and use these features in conjunction with the values used in this study19-22. The accuracy boost will come mainly from how the key features are selected, which should depend on the experimenter’s knowledge of how AD affects which specific areas of the brain in particular at the early onset stage.

1. **CONCLUSION**

In studying the demographic, clinical, and neuromorphic data from MRI scans, it is entirely possible to build an accurate model using machine learning classifiers. The process described in this paper can be used to implement an auxiliary diagnostic tool to be used by physicians in dealing with early detection of AD. The best approach would be to refine the use of kernel SVMs.

Going forward, it would result in an improvement in accuracy if more pertinent key features could be extracted from MRI scans, and more data from other subjects could be obtained and merged with the currently available data from OASIS.

**ACKNOWLEDGEMENTS**

The author would like to thank OASIS for publishing their findings and making them freely available, without which none of this would have been possible. The author recognizes the following grant numbers: P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584

The author would also like to thank the people who developed Scikit-learn and are doing such a great job of maintaining it an easy to use state-of-the-art analytic tool23.

**REFERENCES**

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement. 2007 Jul;3(3):186-91
2. Li T, Braunstein KE, Zhang J, Lau A, Sibener L, Deeble C, Wong PC. The neuritic plaque facilitates pathological conversion of tau in an Alzheimer's disease mouse model. Nat Commun. 2016 Jul 4;7:12082
3. Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Jr, Philip Scheltens, and Paul M. Thompson. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol. 2010 Feb; 6(2): 67–77.
4. Linda K McEvoy and James B Brewer. Quantitative structural MRI for early detection of Alzheimer’s disease. Expert Rev Neurother. 2010 Nov; 10(11): 1675–1688.
5. Buckner RL, Head D, Parker J, Fotenos AF, Marcus D, Morris JC, Snyder AZ. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. Neuroimage. 2004 Oct;23(2):724-38.
6. Lenore Kurlowicz, PhD, RN, CS; Meredith Wallace, RN, MSN, PhD. The Mini-Mental State Examination (MMSE). Journal of Gerontological Nursing. May 1999 - Volume 25 · Issue 5: 8-9.
7. Arevalo-Rodriguez I, Smailagic N, Roqué I Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, Pedraza OL, Bonfill Cosp X, Cullum S. Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev. 2015 Mar 5;(3):CD010783
8. Karp A, Kåreholt I, Qiu C, Bellander T, Winblad B, Fratiglioni L. Relation of education and occupation-based socioeconomic status to incident Alzheimer's disease. Am J Epidemiol. 2004 Jan 15;159(2):175-83.
9. Emily Schoenhofen Sharp, M.A. and Margaret Gatz, Ph.D. The Relationship between Education and Dementia An Updated Systematic Review. Alzheimer Dis Assoc Disord. 2011 Oct; 25(4): 289–304.
10. The Clinical Dementia Rating (CDR) <http://alzheimer.wustl.edu/cdr/cdr.htm>
11. Kim JW, Byun MS, Sohn BK, Yi D, Seo EH, Choe YM, Kim SG, Choi HJ, Lee JH, Chee IS, Woo JI, Lee DY. Clinical Dementia Rating Orientation Score as an Excellent Predictor of the Progression to Alzheimer's Disease in Mild Cognitive Impairment. Psychiatry Investig. 2017 Jul;14(4):420-426.
12. Di Deco J, González AM, Díaz J, Mato V, García-Frank D, Álvarez-Linera J, Frank A, Hernández-Tamames JA. Machine learning and social network analysis applied to Alzheimer's disease biomarkers. Curr Top Med Chem. 2013;13(5):652-62.
13. <http://www.oasis-brains.org/pdf/oasis_cross-sectional_facts.pdf>
14. Tom M. Mitchell. Machine Learning. McGraw-Hill(1997). ISBN: 007042877
15. Ibid.
16. Ibid.
17. Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer(2007). ISBN-10: 0-387-31073-8. ISBN-13: 978-0387-31073-2.
18. Y. Zhang, S. Wang, and Z. Dong. Classification of Alzheimer Disease Based on Structural Magnetic Resonance Imaging by Kernel Support Vector Machine Decision Tree. Progress In Electromagnetics Research. 2014. Vol. 144, 171-184.
19. Yudong Zhang and Shuihua Wang. Detection of Alzheimer’s disease by displacement field and machine learning. PeerJ. 2015; 3: e1251.
20. Saima Farhan, Muhammad Abuzar Fahiem, and Huma Tauseef. An Ensemble-of-Classifiers Based Approach for Early Diagnosis of Alzheimer’s Disease: Classification Using Structural Features of Brain Images. Computational and Mathematical Methods in Medicine. Volume 2014, Article ID 862307.
21. Abdalla R. Gad, N. M. Hussein Hassan, Rania A. Abul Seoud, Tamer M. Nassef. Automatic Machine Learning Classification of Alzheimer's Disease Based on Selected Slices from 3D Magnetic Resonance Imagining. International Journal of Biomedical Science and Engineering. Volume 4, Issue 6, December 2016, Pages: 50-54.
22. Chaiyaporn Mutsalklisana, Kishore Mohan, Akshaya Nagarajan, Priyanka Mishra. Machine Learning Approach to Improve Prediction Accuracy of Alzheimer’s Disease. ASRJETS, 2016. Vol 20, No 1.
23. [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

1. In the original dataset there were three subjects with CDR rating of 2, which posed problems with k-fold CV, so were discarded in the final version of the study. [↑](#footnote-ref-1)