Journal of Neuroscience Methods 375 (2022) 109582

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Addressing the missing data challenge in multi-modal datasets for the diagnosis of Alzheimer's disease

Maryamossadat Aghili *, Solale Tabarestani *, Malek Adjouadi *

Center for Advanced Technology and Education Department of Electrical and Computer Engineering Florida International University Miami, FL, USA

ARTICLE INFO

Keywords:
Alzheimer's Disease
Gradient Boosting (GB)
Support Vector Machine (SVM)
Random Forest (RF)
Soft Impute
SVD Impute
Weighted K-nearest neighbors (KNN impute)
ADNI data
Multiclass classification
Multimodal data

ABSTRACT

Background: One of the challenges facing accurate diagnosis and prognosis of Alzheimer's disease, beyond identifying the subtle changes that define its early onset, is the scarcity of sufficient data compounded by the missing data challenge. Although there are many participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, many of the observations have a lot of missing features which often leads to the exclusion of potentially valuable data points in many ongoing experiments, especially in longitudinal studies.

New methods: Motivated by the necessity of examining all participants, even those with missing tests or imaging modalities, this study draws attention to the Gradient Boosting (GB) algorithm which has an inherent capability of addressing missing values. The four groups considered include: Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI) and Alzheimer's Disease (AD). Prior to applying state of the art classifiers such as Support Vector Machine (SVM) and Random Forest (RF), the impact of imputing (i.e., replacing) data in common datasets with numerical techniques has been investigated and compared with the GB algorithm. Empirical evaluations show that the GB performance is highly resilient to missing values in comparison to SVM and RF algorithms. These latter algorithms can however be improved when coupled with more sophisticated imputation technique such as soft-impute or K-Nearest Neighbors (KNN) algorithm assuming low extent of data incompleteness.

Results: The classification accuracy has been improved by up to 3% in the multiclass classification of all four classes of subjects when all the samples including the incomplete ones are considered during the model generation and testing phases.

Comparison with existing methods: Unlike other methods, the proposed approach addresses the challenging multiclass classification of the ADNI dataset in the presence of different levels of missing data points. It also provides a comparative study on effects of existing imputation techniques on a block-wise missing data. Results of the proposed method are validated against gold standard methods used for AD classification.

1. Introduction

Alzheimer's disease (AD) is one of the most prevailing causes of dementia, which leads to memory loss and other cognitive impairments. This disease is accountable for 60–80% of dementia cases (Barnes and Yaffe, 2011). Accurate AD diagnosis and prognosis is of critical importance, especially for detecting the early stages of the disease through more precise delineation of the Early Mild Cognitive Impairment (EMCI) group from the cognitively normal (CN) control group, and for discriminating possible converter mild cognitive impaired patients from non-converter subjects (Cuingnet et al., 2011; Petersen and Morris, 2005).

Early diagnosis allows for the planning of early treatment and therapeutic interventions, and plays a significant role in providing subject-specific care, predicting disease progression, and gauging the rate of decline and severity of impairment (Moradi et al., 2015; Izquierdo et al., 2017; Dickerson et al., 2009). The more recent studies have devoted great efforts for the early detection of AD by developing interesting algorithms for delineating the prodromal stage of mild cognitive impairment (MCI) with varied and competing classification results (Suk et al., 2014; Landau et al., 2010; Jagust, 2018). Most researchers agree that the current accuracy achieved, especially for classifying the challenging EMCI vs. CN groups remains below an acceptable standard for the medical field considering the irreversible nature of the disease. This

* Corresponding authors.
E-mail addresses: maghili001@fiu.edu (M. Aghili), Staba006@fiu.edu (S. Tabarestani), adjouadi@fiu.edu (M. Adjouadi).

https://doi.org/10.1016/j.jneumeth.2022.109582
Received 18 April 2021; Received in revised form 22 March 2022; Accepted 23 March 2022
Available online 26 March 2022

error in classification is of course further influenced in the negative when performing the more realistic multiclass classification (assuming all groups as it should be) rather than a presumed binary classification (when only two classes are considered at a time). Moreover, when dealing with multimodal and multiclass classification, researchers often contend with the missing data challenge, especially when longitudinal studies are considered. Lack of sufficient data with complete samples for all the subjects considered in a study whether cross-sectional or longitudinal is an inherent problem of any clinical trial. This challenge of missing data continues to hinder the needed progress for understanding this challenging and complex brain disorder (Jagust, 2013).

In the medical field, incomplete samples in longitudinal studies are frequent. This is largely due to patients who miss taking some of the tests at some different timepoints of a study. Generally, missing values occur for a variety of reasons, including subjects that miss appointments, subjects that completely drop out from the study, budget limitation or when dealing with data with insufficient or incompatible resolutions or experience image corruption, etc. Troyanskaya et al. (2001); Lo and Jagust (2012). Many algorithms simply discard subjects with missing modalities from further consideration or, in the simplest case, they just replace the missing data with zero values or with a mean average of the attribute, which still results in a loss of valuable information. Accuracy in AD diagnosis and prognosis could be improved if the missing parameters can be more precisely estimated from the rest of the available data through reliable machine learning techniques, rather than through standard substitution techniques (Belger et al., 2016). Added attention is needed when different data modalities often have nonlinear and complicated correlations, which impedes the prospects for correct estimation.

These challenging issues have led to a new line of research that focuses on developing more realistic and more sophisticated techniques to resolve experimental issues involving incomplete samples. This line of research is generally divided into two main approaches: the first approach attempts to synthesize missing modalities from the remaining ones with the help of various techniques that include maximum mean discrepancy based multiple kernel learning (Zhu et al., 2017), cascaded residual autoencoder (Tran et al., 2017), 3D convolutional neural networks (Payan and Montana, 2015) and generative adversarial networks (GAN) (Nie et al., 2017; Xiang et al., 2018). Regarding the application of GAN in medical imaging, Cohen and his colleagues have pointed out that synthesized medical images may result in misdiagnosis due to the distribution matching losses that arise from the process of matching an image in the input domain to an image in the target domain while preserving the source distribution (Cohen et al., 2018). The second approach attempts to impute missing values by applying various numerical techniques such as simple Mean substitution,[1] Mode and K-Nearest Neighbor (KNN) impute (Campos et al., 2015; Luengo et al., 2012; Huang et al., 2016). Authors in (Xiang et al., 2018; Ritter et al., 2015) extracted a complete subset of features from the actual dataset and synthesized the missing values randomly to analyze the power of some imputation methods, but they have not tested the algorithms on different patterns of missing values in real incomplete datasets, which may actually have completely different patterns from those that were randomly synthesized. They also overlooked the fact that some of the proposed imputation methods assume that the data have a Gaussian distribution, which may not be the case for every dataset. Moreover, some of these approaches do not address the block-wise missing patterns of data in the relatively small dataset size of the AD group. When the data is multimodal in nature acquired through MRI, PET, CSF, and cognitive scores, to name a few, each modality creates multiple features in each sample. When a modality is missing for a subject then none of those features from that single modality will be available for that sample leading to a missing block of information called block-wise missing pattern.

Therefore, to the best of our knowledge, none of the research studies so far have done a comparative study on effects of existing imputation techniques on a block-wise missing dataset of Alzheimer while incorporating a huge sample size from various modalities to check the effects of large size data on imputation tasks. As an additional task, we have also considered the challenging multiclass classification of the ADNI dataset in the presence of a high number of missing points. Moreover, there are several new imputation techniques which have never been deeply studied within this scope of work.

Considering the importance of the early detection of the prodromal stage of AD, the first objective of this paper is to analyze the classification power of Gradient Boosting (GB) technique on a four-way classification. The four groups included Cognitively Normal controls (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI) and Alzheimer Disease (AD) acquired from a large multimodal heterogeneous dataset pulled from various cites with missing data, which include ADNI1, ADNIII and ADNIGO. The recent release of ADNI data which discriminate early, and late MCI patients motivated us to focus on multiclass classification between the four groups of subjects rather than using binary classification of two groups of subjects at a time. The assumption here is that binary classification lacks the generalization power when introducing new sample data with no prior diagnosis label. The challenge of discriminating the EMCI group from LMCI has not yet been well studied due mainly to the absence of adequate data for those two classes. The second objective of this paper is to represent the classification potential of GB and its potential to handle incomplete data sets. Experimental evaluations show that SVM is unable to work with incomplete sample data, GB is capable of handling missing values with no need for any additional preprocessing.

We also describe the performance dependency of the various state-of-the-art imputation techniques on the patterns of missing data. For this purpose, we investigated the performance of a group of imputation techniques on two separate sets of synthesized incomplete data with random-wise missing values and real incomplete data with block-wise missing values. Results reveal the shortcomings of imputation techniques in the real case of block-wise missing data estimation. Despite few papers that attempted to proceed in this direction (Campos et al., 2015; Jiang et al., 2016), to the best of our knowledge, this work is the first one that provides an extensive comparative study over real, incomplete heterogeneous multimodal dataset of Alzheimer with the four groups: CN, EMCI, LMCI, and AD.

The remainder of this paper is organized as follows: Section II describes the dataset and the preprocessing steps that were undertaken. Section III defines the methods that have been investigated and implemented in this study. Section IV provides the experimental results and related analyses. Finally, Section V closes with the discussion and conclusion.

2. Dataset and preprocessing

Data used in the preparation of this article were obtained from the Alzheimer Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and Alzheimer's Disease (Landau et al., 2010).

ADNI data is processed with a standard pipeline resulting in a large matrix of patients and their test measurements. Patients are arranged in rows and each test result is ordered as a column. In this paper, we used various groups of biomarkers including CSF, MRI, PET, DTI, Genetics, and neuropsychological tests, which are derived from ADNI database.

[1] Since data is normalized around the center in this study, mean substitution in this case is the same as zero fill.

Table 1
Biomarkers Used in this Study.

Source	Features
Cognitive tests	Rey Auditory Verbal Learning Test (RAVLT Immediate, RAVLT Learning, RAVLT Forgetting, RAVLT Perc Forgetting), Functional Activities Questionnaires (FAQ), Everyday Cognition (Ecog) scales: (EcogPtMem, EcogPtLang, EcogPtVisspat, EcogPtPlan, EcogPtOrgan, EcogPtDivatt, EcogPtTotal, EcogSPMem, EcogSPLang, EcogSPVisspat, EcogSPPlan, EcogSPOrgan, EcogSPDivatt, and EcogSPTotal)
MRI	Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, MidTemp, ICV, FLDSTRENG, FSVERSION
PET	FDG, PIB amyloid, AV45 amyloid, CDRSB
Genetic	APOE4
Demographic	AGE, Gender, Education
CSF	Ab1, t-tau, p-tau

The detailed list of biomarkers is provided in Table 1. Diagnosis labels consist of Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI) and Alzheimer's Disease (AD).

The dataset considered for this study consists of 1627 subjects, among them are 413 CN, 312 EMCI, 565 LMCI and 337 AD subjects, which have been examined for up to an 11-year period with visits scheduled every six months. However, inherent to all longitudinal studies, many of these subjects tend to miss few to many of these visits due to dropout, relocation of patient, scheduling issues or health-related reasons. Some of them have only a couple of visits or time points throughout the duration of the study. A comprehensive set of 41 biomarkers were selected at the end, as indicated in Table 1.

Imputation techniques, like multiclass classification or prediction, tend to perform better when relatively large datasets are available, providing more samples for training the model in search of an optimal classifier. We considered each patient-visit to be a separate sample to augment the data size, help the imputation process and augment the prospects for establishing or modeling an optimal classifier that generalizes better. However, to avoid data leakage, we carefully split the data based on patient ID and made sure that samples of the same patient do not appear in both training and testing phases. It is worth mentioning that even if the subject status does not change between successive time points, subject-sample vector at various time points tends to be slightly different from each other due to changes in imaging and other performed tests.

In contrast to most studies that handle the problem of missing data, either by excluding the patients with incomplete test results or by restricting the study to a single modality, we tried to solve the problem by introducing an algorithm which can handle the missing values naturally. In a parallel experiment, we integrate an imputation stage to two other classifiers of RF and SVM to check if this adds any robustness to the classification algorithms. Furthermore, in contrast to most studies which work on a dataset from a specific single source, current ADNIMERGE dataset is pulled from multiple cites which adds heterogeneity to the data and makes the classification process even more challenging. Pre-analysis of the dataset indicates that based on all the biomarkers considered in this study, there was not a single record that did not have one or more measurements missing. The number of missing values altogether for all data samples and throughout all the biomarkers is equivalent to 46% of the entire dataset. This clearly highlights the extent of the missing data challenge researchers face when considering longitudinal studies. It also places research groups in the predicament of choosing only those datasets with the complete set of measurements, ending up with a much smaller dataset and perhaps less statistically meaningful. The best option is to consider the entire data set and to find effective ways to impute the missing data with the aim to preserve the statistical and clinical meaningfulness of the data that took so much effort and so many years to acquire.

Since feature normalization is required for many algorithms, especially in the cases of SVM and K-nearest neighbors, datasets are centered and normalized. As there are many missing values in the dataset, before normalization, missing values are masked. Hence, the convergence time was reduced dramatically, and classification accuracy is improved, especially when using SVM.

3. Experimental methods

We investigated the outcomes of weighted K-Nearest Neighbors (KNN) (Zhang, 2012) Singular Value Decomposition (SVD) based method (Golub and Reinsch, 1971), Soft Impute (Mazumder et al., 2010), Matrix Factorization (Paatero and Tapper, 1994) and Mean average, combined with three state of the art classification algorithms; Support Vector Machines (Suykens and Vandewalle, 1999), Random Forest (Liaw and Wiener, 2002), and Gradient Boosting (Ogutu et al., 2011; Friedman, 2001; Chen and Guestrin, 2016). To analyze the methods precisely, we have repeated the experiments with varying percentages of missing values. Hyper parameters were carefully adjusted by performing an exhaustive grid search to reach the best performing classifier. In this section, we briefly overview the methods that have been exploited in this study. All the methods are implemented using Fancy-impute libraries as detailed in section IV.

The K-Nearest Neighbors imputation method selects subjects with similar feature sets to the subject that has missing values. For example, if a sample S has a missing value in feature Q, this method would select all the subjects which are most similar to S and have that feature Q. It gives a weight to each retrieved sample based on the degree of similarity and then calculates the weighted average as the estimated value for the missing target in sample S. For the similarity measure, various metrics can be utilized such as Pearson correlation, Euclidean distance, and variance minimization. In our study we used the Euclidean distance as the similarity measure of the data (Zhang, 2012).

Matrix Factorization method was first introduced in (Paatero and Tapper, 1994) and since then it has been used in many applications such as collaborative filtering and missing value imputations. This technique attempts to split the original large matrix of $X \in R^{n*m}$, in which n is the number of subjects and m is the number of features, into two matrix components of smaller dimensions as a function of a k factor, $W \in R^{n*k}$ and $H \in R^{k*m}$. Since the original matrix of samples and features has a lot of missing values, the sparsity constraint is imposed on matrix H which results in the following minimized formulation:

$$W, H \min \frac{1}{2}\left[\left\|X - WH^T\right\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2\right] \quad (1)$$

subject to $W, H \geq 0$

with α and β being the regularizing constants and $\|\|_F^2$ defining the Frobenius Norm. To reach the global minima, the mentioned minimization problem is solved using gradient descent (Paatero and Tapper, 1994).

The Singular Value Decomposition (SVD) method has been proposed in (Suykens and Vandewalle, 1999), which is another approach for estimating the missing data iteratively. Assume that X is a set of observed elements and X^r as a subset of X. SVD-impute applies singular value decomposition of matrix X to get orthonormal patterns of U and V. The approximation of X^r can then be derived by a linear combination of these patterns through $J_J D_J V_J^r$ where J_J, D_J and V_J^r are orthogonal. Then, the SVD imputation of any matrix X can be implied by solving the following problem:

$$Min\left\|X - m_i^r - U_J D_J V_J^r\right\| \quad (2)$$

where m_i^r is the mean of the i^{th} row and $\|\|$ is a sum of squared values of all non-missing elements. In this method, we start the procedure by substituting the missing values in X by the means of all non-missing values in each row. Then (2) will be solved for a new set of matrices of U, V and D which produces a new approximation of X. This step will

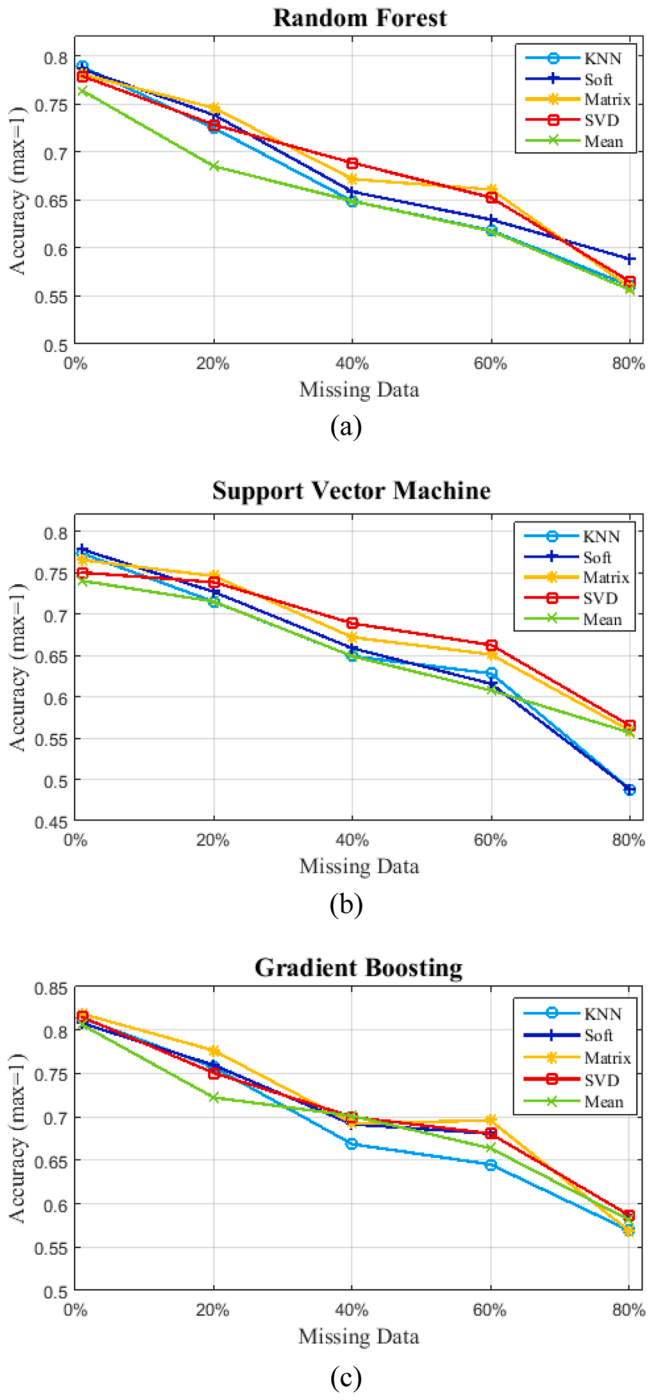


Fig. 1. Effects of imputation methods coupled with three classifiers at different degrees of random missing values (a) Random Forest (b) Support Vector Machine (c) Gradient Boosting on 4-way multiclass classification (CN, EMCI, LMCI, AD) of the subjects.

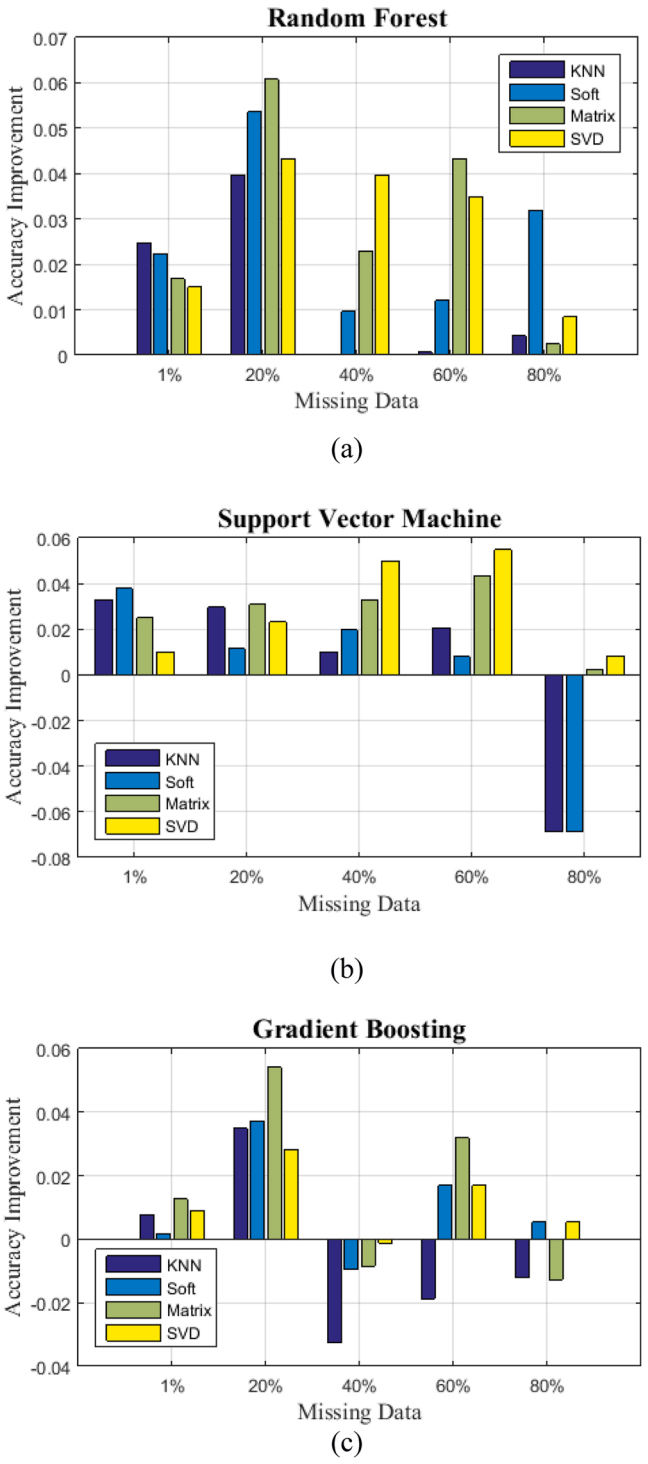


Fig. 2. The improvement achieved by cascading different imputation techniques with classification methods (a) Random Forest (b) Support Vector Machine (c) Gradient Boosting on 4-way classification of subjects.

be repeated until the difference between the $X_{i+1}$and X_i meets the optimal stopping criteria (Golub and Reinsch, 1971).

Soft Impute has been proposed in (Mazumder et al., 2010) as a more efficient algorithm than the original iterative SVD which addresses the high computational cost of iterative SVD for large matrices. However, it computes a low-rank SVD of a dense matrix repetitively. This allows the regularization path of solutions to be computed efficiently on a grid of regularization parameters. Rank reduction and shrinkage is performed simultaneously in soft impute in a single operation. More precisely, this algorithm solves Eq. (3) to deduce and replace the missing values. Then the SVD imputation of any matrix X can be implied by solving this problem:

$$\min_{Z} \frac{1}{2} \quad \|W - Z\|_F^2 + \quad \lambda\|Z\| \tag{3}$$

where λ is a regularization parameter. This algorithm initializes the missing values with zero and keeps track of the old Z and replaces the $Z^{new}$with $S_{\lambda_k}\left(P_{\Omega}(X) + P_{\Omega}^{\perp}\left(Z^{old}\right)\right)$, until it hits the exit or stop criteria

Table 2
Correlation coefficient of datapoints in original data versus predicted via different imputation techniques with varying percentage of random missing values.

Method / Missing %	1%	20%	40%	60%	80%
KNN	0.999	0.998	0.995	0.993	0.988
Soft Impute	0.999	0.998	0.986	0.888	0.727
Matrix Factorization	0.999	0.999	0.997	0.994	0.993
SVD	0.999	0.980	0.948	0.786	0.589
Zero fill	0.998	0.838	0.751	0.634	0.310

defined below:

$$\frac{\left\|Z^{new} - Z^{old}\right\|_F^2}{\left\|Z^{old}\right\|_F^2} < \varepsilon \tag{4}$$

$$\left\{P_\Omega(X)(i,j) = \begin{cases} X_{ij}, \& if & (i,j) \in \Omega \\ 0 \quad , \& if & (i,j) \notin \Omega \end{cases}\right\} \tag{5}$$

$P_\Omega(X)$ of dimension m*n is a projection of matrix X onto the observed entries. $P_\Omega^\perp(X)$ is a complementary projection such that $P_\Omega(X) + P_\Omega^\perp(X) = X$. The above low-rank optimization models are usually used for collaborative filtering, nonetheless they have application in other domains such as missing data imputation, clustering and data retrieval.

Support Vector Machines (SVM) remain a robust statistical method first introduced in the early 1990 s as a nonlinear solution for regression and classification (Suykens and Vandewalle, 1999). This technique has been proven to have superior performance in addressing various problems due to its generalization abilities, robustness against noise and other forms of interference, and its computational efficiency is comparable to several other methods. Support vector machines separate two or more classes by finding an optimal hyperplane with a maximized margin known as support vectors. Multi-class SVM problems can be solved by decomposition into a predefined number of binary problems. Two known approaches are one-versus-rest and one-versus-one. One--versus-rest classifiers are composed of k separate binary classifiers in which each classifier will be trained using the data of its own class with a positive outcome and the data from all other classes as negative outcome. One-versus-one approach is composed of all pairwise individual classifiers where each test example will be fed into all individual classifiers and the data is assigned to the class which yields the highest winning score (Ogutu et al., 2011).

Random Forest (RF) is a type of supervised machine learning algorithm which is an ensemble of multiple decision trees. For each tree in the forest a bootstrap sample of data is taken to create various input datasets so that each tree will be fit with a different set of samples. Then the data will be split based on a selection of random variables. The best split will iteratively be selected based on the impurity measure. The whole process will be repeated in building several decision trees to complete the random forest model. Each new data point will be fed iteratively into all generated trees and their outcome will be averaged to form the final prediction of the random forest (Liaw and Wiener, 2002).

Gradient Boosting (GB) is a powerful supervised machine learning technique commonly used to solve regression, multiclass classification, and ranking problems. This technique has a sequence of weak tree learners which are trained to fit a given model K such that each learner will improve the prediction accuracy of the previous one by minimizing the multiclass logistic likelihood J between the pseudo residuals using the following formula:

$$J = \sum_i L(y_i, K(x_i)) \tag{6}$$

In which y_i is the target value and $K(x_i)$ is the value obtained from the predicted model. GB is robust to redundant data and has the inherent ability to handle missing data. During the training phase, GB computes the optimal split direction for every feature, therefore it decides if missing values should go to either right or left node of the tree to minimize the loss function. Hence, we were interested in using Gradient Boosting in our study to examine and test the embedded imputation strength of this algorithm (Ogutu et al., 2011; Friedman, 2001; Chen and Guestrin, 2016) against the proposed cascaded imputation-classification method.

4. Evaluation on subjects

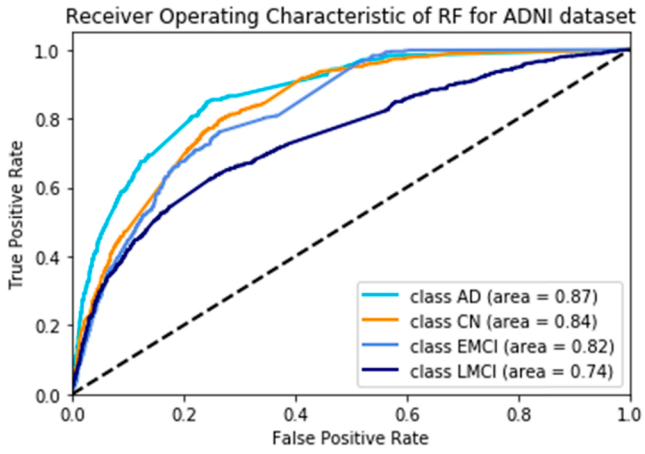
The experiments conducted proceed through multiple steps. At the first step, 70% of the data is randomly selected for training, 10% used as the validation set, and the remaining 20% is used in the testing phase. As one subject may be tested multiple times and appears under various sample ID, the data split has been performed based on the unique subject ID instead of the sample ID. In this way, the chance of visiting similar samples of the same patient in both the training and testing phases is removed. The data was normalized by subtracting the mean value and then dividing by the standard deviation prior to imputation. A mask was generated to cover the Not Available (NA) or missing values when needed.

The next step involved estimating the missing data using different imputation techniques including KNN impute, iterative SVD, Matrix Factorization, Soft Impute and Mean averaging. After that, the classification of subjects is performed. We implemented the code in Python using *Scikit-learn* module for machine learning (Pedregosa et al., 2011) and Fancy-impute libraries. While other classifiers were more robust when non-normalized data were used, SVM accuracy improved dramatically with normalization.

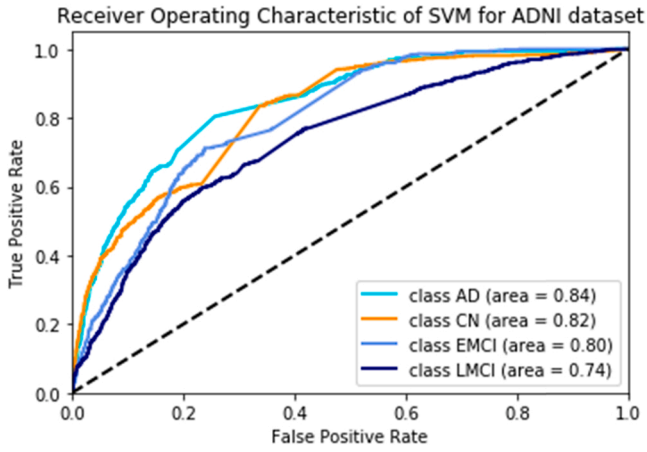
We excluded the diagnosis labels (CN, EMCI, LMCI, AD), and some of the highly correlated cognitive test scores such as the Mini Mental State Examination (MMSE), Clinical Dementia Rating scale (CDR) as well as the Alzheimer's Disease Assessment Scale-Cog (ADAS-Cog) from the training phase to avoid introducing any bias in the the results. Moreover, comparative assessments to other studies will be fair only if similar features/modalities and similar datasets are used. In this study, imputation has been done across training, validation, and testing data separately prior to classification. Each classifier has been adjusted through an exhaustive grid search with cross validation to achieve optimal accuracy. Tuning parameters for the SVM method consisted of a Gaussian-based radial basis function (RBF) kernel with Gamma and C parameters set to 0.0001 and 100, respectively. For RF, the maximum number of features at each node was set to 10, the minimum number of samples required in each leaf was set to 3, and the minimum number of samples

Table 3
Comparison of (a) Random Forest (b) Support Vector Machine (c) Gradient Boosting coupled with five imputation techniques* .

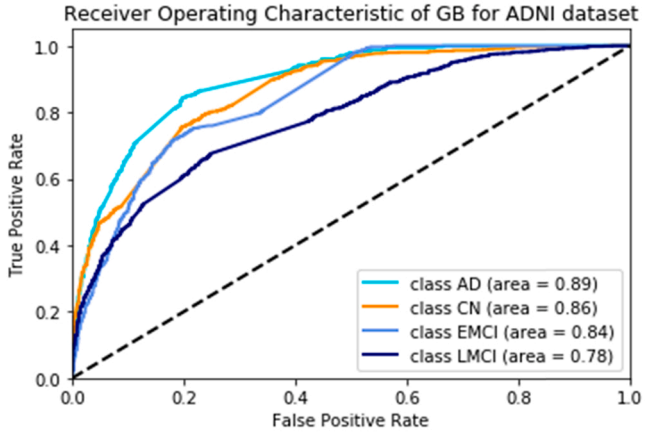
Classifier	*Gradient Boosting*			*Support Vector Machine*			*Random Forest*		
	Accuracy	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*
KNN	59.21	60.50	57.10	58.89	51.80	55.29	53.65	56.82	51.61
Soft Impute	65.03	63.12	63.00	58.88	54.61	57.91	57.54	58.10	55.65
Matrix Factorization	62.10	60.65	59.00	57.48	48.92	56.34	55.61	57.02	53.10
Iterative SVD	62.20	62.27	62.20	58.83	60.49	56.88	58.18	55.91	56.41
Mean	**62.12**	**63.32**	**62.20**	57.70	52.58	55.47	60.56	60.59	59.65



(a)



(b)



(c)

Fig. 3. Comparison of ROC curve for three classifiers (a) Random Forest with mean imputation (b) Support Vector Machine with mean imputation (c) Gradient Boosting without imputation technique in a 4-way classification of subjects.

required to split an internal node is set to 2, with a Gini index for criteria of quality split. For the GB method, the maximum depth of individual regression estimators was set at 2, number of features at 25, subsample used for fitting learner at 14, minimum number of samples at 10, and number of boosting stages at 28.

To attain a robust performance prediction, we repeated all experiments over 30 trials and the metrics across all trials have been averaged. Besides providing accuracy, we also provide performance evaluation metrics that include precision, recall, and Receiver Operating Characteristic curve (ROC) which is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. In the following section, the experiment is explained with more details.

4.1. *Synthetic missing data handling*

To have a clear understanding of the effect of imputation techniques on diverse patterns of missing values, performance of each technique is evaluated on a complete extracted version of the original dataset where only observations with no missing values were retained. Subsequently, we randomly deleted 1%, 20%, 40%, 60%, and 80% of the data to investigate the best combination of classification-imputation pairs that achieves an optimal classification accuracy. Experimental evaluations prove that the SVM and RF classifiers, when coupled with an imputation method like Matrix Factorization, Soft Impute or KNN produces the highest accuracy in multiclass classification as compared to mean substitution in almost all percentages of missing values for up to 80%. From the results shown in Fig. 1, there is enough evidence that selecting an appropriate imputation technique can improve the accuracy of SVM and RF in case of missing data with a random pattern. However, GB achieves relatively equivalent result with or without the imputation techniques owing to its innate ability for handling missing data.

Fig. 1. also demonstrates that from low to middle percentage of random missing values, GB exhibits the highest classification accuracy, regardless of the imputation technique used. To illustrate the effectiveness of different imputation techniques in the presence of various amounts of synthesized missing data, we calculated the accuracy improvement for each classifier as shown in Fig. 2. For a low percentage of synthesized random missing values, the highest improvement in accuracy is achieved by RF coupled with Matrix Factorization technique as shown in Fig. 2. (a) which happened at 20% of missing data. Additionally, this experiment shows that for a high percentage of missing values, none of the imputation techniques can estimate patterns of missing data correctly.

All these experiments have been repeated 30 times and the standard deviation in accuracy improvement of RF and SVM coupled with imputation techniques over the 30 random runs was mostly between 2% and 4% over the different percentages of missing values. Higher standard deviation values of 5% and 6% resulted with the GB method which seemed not to gain in accuracy when coupled with any imputation technique. This indicates that GB may not benefit as well from imputation techniques, since it intrinsically addresses the missing values in the way it is originally conceived.

The coefficient of correlation between the original dataset and what is generated by different imputation techniques have been calculated to give a fair ground for comparing the performance of the imputation techniques when they address the various percentages of missing data. It can be observed from the results shown in Table 2 that as the percentage of the missing values increases, the covariance coefficient declines, while KNN and Matrix Factorization have shown the highest correlation coefficient overall.

4.2. *Original Missing Data Handling*

We repeated our experiments on the original incomplete dataset, where almost 40% of the data is missing and the pattern considered in this case is not random but is assumed block-wise missing (Lo and Jagust, 2012). Considering the measurements summarized in Table 3, it can be observed that GB method yields the best results over all combinations of classifiers and type of value substitution for GB can be observed through the ROC curves in Fig. 3. The ROC curves of the classifiers represent the difficulty in delineating the four classes (CN,

Table 4
Binary Classification of the Control Normal vs Early Mild Cognitive Impairment (CN vs EMCI) * .

Classifier	*Gradient Boosting*			*Support Vector Machine*			*Random Forest*		
Imputation tech	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***
KNN	70.28	70.54	67.20	78.18	77.39	77.01	72.01	70.14	70.11
Soft Impute	79.07	76.79	76.27	76.91	75.19	75.15	73.71	72.39	72.47
Matrix Factorization	80.23	77.21	77.21	76.54	73.45	73.44	75.85	73.54	73.28
SVD	80.19	79.00	79.51	76.08	74.54	72.88	74.97	73.72	73.94
Mean	**80.93**	**79.67**	**79.9**	76.96	75.93	76.22	75.56	73.91	74.33

Table 5
Binary Classification of the Early vs Late Mild Cognitive Impairment (EMCI vs LMCI) * .

Classifier	*Gradient Boosting*			*Support Vector Machine*			*Random Forest*		
Imputation tech	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***
KNN	83.19	81.32	81.98	83.13	79.11	80.34	79.98	78.45	79.28
Soft Impute	82.75	79.53	80.71	82.42	81.72	82.34	83.56	80.67	81.63
Matrix Factorization	85.12	81.14	82.16	77.35	76.66	77.44	81.55	79.73	79.54
SVD	85.81	81.12	82.33	79.12	77.65	79.29	81.61	80.13	80.26
Mean	**85.84**	**82.22**	**82.40**	81.38	81.20	82.4	83.24	80.74	81.74

Table 6
Binary Classification of the Late Cognitive Impairment Vs Alzheimer (LMCI vs AD) * .

Classifier	*Gradient Boosting*			*Support Vector Machine*			*Random Forest*		
Imputation tech	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***
KNN	73.05	67.24	69.53	65.55	46.44	64.07	73.92	70.78	72.2
Soft Impute	74.84	71.08	72.42	69.92	47.79	68.40	73.45	69.98	71.4
Matrix Factorization	73.84	70.87	71.81	66.25	43.58	63.97	72.32	68.31	70.2
SVD	74.23	70.42	71.34	69.32	46.75	69.47	74.63	72.46	72.6
Mean	**75.23**	**73.16**	**73.35**	70.12	44.35	66.87	74.59	71.67	72.29

Table 7
Binary Classification of the Control Normal Vs Alzheimer (CN vs AD) * .

Classifier	*Gradient Boosting*			*Support Vector Machine*			*Random Forest*		
Imputation tech	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***
KNN	93.30	91.54	91.14	91.82	90.78	90.24	92.77	90.52	90.14
Soft Impute	91.67	91.36	91.07	90.89	91.02	91.08	91.28	89.61	89.83
Matrix Factorization	91.41	91.32	91.34	91.84	90.84	90.18	90.74	90.74	90.85
SVD	90.90	90.88	90.29	90.57	90.61	90.75	89.81	89.15	89.18
Mean	**93.40**	**92.37**	**91.44**	91.85	90.11	89.54	92.71	90.21	90.12

EMCI, LMCI, AD) in a real comprehensive dataset. Among all, GB records the highest AUC for all the classes AUC = 0.89, 0.86, 0.84 and 0.78 for AD, CN, EMCI and LMCI respectively while RF with AUC = 0.87, 0.84, 0.82, 0.74 has the second place and SVM with 0.84, 0.82, 0.80, 0.74 shows a somewhat lower performance for all the classes. Fig. 3. reveals that EMCI and LMCI separation is the most difficult task for all the three classifiers.

Based on the missing data patterns and quantity of the missing data, imputation-classification pairing can perform better than simple mean value substitution, but this improvement highly depends on the distribution of the data. Hence, these investigations reveal that none of the state-of-the-art imputation techniques could address block-wise missing data.

To emphasize the difficulty of multiclass classification, the accuracy of binary classification, which has been a focus of AD related research for many years (Xiang et al., 2013; Gray et al., 2013), is provided in Tables 3–7. These results highlight that even though classification of subjects between two classes at a time provides higher accuracy, F-score, precision and recall in almost all the cases (AD vs CN, CN vs EMCI, EMCI vs LMCI, and LMCI vs AD), these types of classification lack as expected the generalization ability for real-world scenarios when an unseen sample data could belong to any of the four groups of (AD, LMCI, EMCI, CN). Four-way or multiclass classification is hence more desirable and more realistic, but much more challenging especially when dealing with heterogeneous multimodal dataset as the one considered here.

From the results summarized in Tables 3–7, it can be observed that gradient boosting (GB) consistently outperformed the other two algorithms of RF and SVM by at least a 2% improvement in accuracy, while maintaining the highest precision and recall scores. Moreover, it is observed that RF and SVM methods could not reach similar performance even if augmented with the most advanced imputation methods. The GB performance also does not change noticeably when paired with these imputation techniques, given that the GB method is designed to address this issue innately.

In our study, although GB performed only slightly better than other methods (2% higher accuracy), it holds perhaps the greatest promise because of its versatility, allowing it to assume simpler, and more interpretable forms, such as component-wise boosting and the ability to incorporate automatic predictor selection. This study also provides evidence that imputation cost in terms of computational overhead is more realistic when the percentage of missing values is under 40% with the pattern of missing data assumed random.

All algorithms evaluated in this study are robust and successful when considering large feature sets. However, SVM works well for smaller

number of observations. RF, on the other hand, is preferable for large non-normalized datasets. SVD and KNN use the correlation structure of the data and KNN uses the Euclidean distance to measure similarity and profile most related observations to estimate the missing values. These approaches will fail to find the most similar profile when it comes to outliers. This flaw can be resolved with scaling or using log over observations. In addition, although the superiority of SVM against other machine learning algorithms in terms of accuracy has been reported in many studies, this study shows that GB achieved higher performance in ADNI dataset with its inherent capability of managing the missing values. RF and GB are also quite robust with respect to collinearity. However, SVM alleviates the multi collinearity problem via regularization, where in RF, it is alleviated via choosing a random subset of features for each tree.

5. Conclusion

In this paper, we presented a comparative study of several methods for the estimation of missing values in the largest heterogeneous dataset pulled from various longitudinal studies and cites. We discussed the difficulty of classification in the inherent presence of missing values in longitudinal studies especially when dealing with a multimodal heterogeneous dataset. Of the different state-of-the-art algorithms implemented in this study, Gradient Boosting algorithm achieved the best performance when dealing with multiclass classification involving all 4 groups (CN, EMCI, LMCI and AD). The GB method has outperformed SVM and Random Forest algorithms. All the classifiers have been coupled with four advanced imputation techniques including KNN impute, Matrix Factorization, SVD, and Soft Impute and they have been utilized to classify the different stages of AD. When coupled with imputation techniques, Random Forest was the most consistent for improving accuracy through all percentages of missing data, followed by SVM up to 60% missing data; but both failed at the 80% and more of missing data. Despite the contribution of the imputation techniques in missing value estimation in data with low percentage of the random missing data, all the algorithms fail to perform well in high levels of missing data. Moreover, in the presence of block-wise missing data patterns, where a particular modality is completely missing for so many subjects, these imputation methods are not as helpful. While many studies so far focused on binary classification of AD, we went further in performing multiclass classification while contending with the missing data challenge inherent to longitudinal studies.

Moreover, we also provide results of the different binary classifications as well for comparative purposes and for estimating the effect of missing data on such binary classification in contrast to multiclass classification. The imbalanced dataset and insufficient samples in each group of subjects imposed a new constraint on the current classification problem. We tried to tackle this issue by incorporating the data samples from longitudinal studies and provided effective ways to augment the dataset. In future work, we are planning on improving the current multiclass classification accuracy with application of newer techniques such as the Optimal Margin Distribution (Zhang and Zhou, 2019) in incomplete datasets, even in the presence of block-wise missing data patterns, and applying new deep learning techniques such as Long Short Term Memory for handling missing data (Aghili et al., 2018; Li et al., 2019).

Acknowledgments

We are grateful for the continued support from the National Science Foundation (NSF) under NSF grants CNS-1920182, CNS-1551221, and CNS-2018611. We also greatly appreciate the support of the 1Florida Alzheimer's Disease Research Center (ADRC) (NIA 1P50AG047266–01A1) and the Ware Foundation. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12–2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California.

References

Aghili, M., Tabarestani, S., Adjouadi, M., & Adeli, E. (2018, September). Predictive modeling of longitudinal data for Alzheimer's Disease Diagnosis Using RNNs. In International Workshop on PRedictive Intelligence In MEdicine (pp. 112–119). Springer, Cham.

Barnes, D.E., Yaffe, K., 2011. The projected effect of risk factor reduction on Alzheimer's disease prevalence. Lancet Neurol. 10 (9), 819–828.

Belger, M., Haro, J.M., Reed, C., Happich, M., Kahle-Wrobleski, K., Argimon, J.M., Wimo, A., 2016. How to deal with missing longitudinal data in cost of illness analysis in Alzheimer's disease—suggestions from the GERAS observational study. BMC Med. Res. Methodol. 16 (1), 1–11.

Campos, S., Pizarro, L., Valle, C., Gray, K.R., Rueckert, D., & Allende, H. (2015, November). Evaluating imputation techniques for missing data in ADNI: a patient classification study. In Ibero-American Congress on Pattern Recognition (pp. 3–10). Springer, Cham.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).

Cohen, J.P., Luck, M., & Honari, S. (2018, September). Distribution matching losses can hallucinate features in medical image translation. In International conference on medical image computing and computer-assisted intervention (pp. 529–536). Springer, Cham.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Alzheimer's Disease Neuroimaging Initiative, 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. neuroimage 56 (2), 766–781.

Dickerson, B.C., Feczko, E., Augustinack, J.C., Pacheco, J., Morris, J.C., Fischl, B., Buckner, R.L., 2009. Differential effects of aging and Alzheimer's disease on medial temporal lobe cortical thickness and surface area. Neurobiol. Aging 30 (3), 432–440.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Golub, G.H., Reinsch, C., 1971. Singular value decomposition and least squares solutions. Linear algebra. Springer,, Berlin, Heidelberg, pp. 134–151.

Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., Alzheimer's Disease Neuroimaging Initiative, 2013. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage 65, 167–175.

Huang, L., Jin, Y., Gao, Y., Thung, K.H., Shen, D., Alzheimer's Disease Neuroimaging Initiative, 2016. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. Neurobiol. Aging 46, 180–191.

Izquierdo, W., Martin, H., Cabrerizo, M., Barreto, A., Andrian, J., Rishe, N.,. & Adjouadi, M. (2017, December). Robust prediction of cognitive test scores in Alzheimer's patients. In 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1–7). IEEE.

Jagust, W., 2013. Vulnerable neural systems and the borderland of brain aging and neurodegeneration. Neuron 77 (2), 219–234.

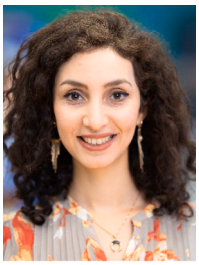
Jagust, W., 2018. Imaging the evolution and pathophysiology of Alzheimer disease. Nat. Rev. Neurosci. 19 (11), 687–700.

Jiang, B., Ma, S., Causey, J., Qiao, L., Hardin, M.P., Bitts, I., Huang, X., 2016. SparRec: An effective matrix completion framework of missing data imputation for GWAS. Sci. Rep. 6 (1), 1–15.

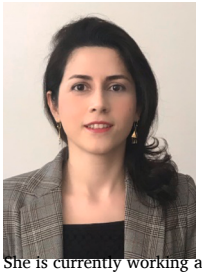
Landau, S.M., Harvey, D., Madison, C.M., Reiman, E.M., Foster, N.L., Aisen, P.S., Jagust, W.J., 2010. Comparing predictors of conversion and decline in mild cognitive impairment. Neurology 75 (3), 230–238.

Li, F., Liu, M., Alzheimer's Disease Neuroimaging Initiative, 2019. A hybrid convolutional and recurrent neural network for hippocampus analysis in Alzheimer's disease. J. Neurosci. Methods 323, 108–118.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. N. 2 (3), 18–22.
Lo, R.Y., Jagust, W.J., 2012. Predicting missing biomarker data in a longitudinal study of Alzheimer disease. Neurology 78 (18), 1376–1382.
Luengo, J., Luengo, S., Herrera, F., 2012. On the choice of the best imputation methods for missing values considering three groups of classification methods, 32 (1), 77–108.
Mazumder, R., Hastie, T., Tibshirani, R., 2010. Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. 11, 2287–2322.
Moradi, Elaheh, et al., 2015. "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects.". Neuroimage 104, 398–412.
Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., & Shen, D. (2017, September). Medical image synthesis with context-aware generative adversarial networks. In International conference on medical image computing and computer-assisted intervention (pp. 417–425). Springer, Cham.
Ogutu, J.O., Piepho, H.P., & Schulz-Streeck, T. (2011, December). A comparison of random forests, boosting and support vector machines for genomic selection. In BMC proceedings (Vol. 5, No. 3, pp. 1–5). BioMed Central.
Paatero, P., Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5 (2), 111–126.
Payan, A., Montana, G., 2015. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv Prepr. arXiv 1502, 02506.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Petersen, R.C., Morris, J.C., 2005. Mild cognitive impairment as a clinical entity and treatment target. Arch. Neurol. 62 (7), 1160–1163.
Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., Haynes, J.D., Alzheimer's Disease Neuroimaging Initiative, 2015. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimer'S. Dement.: Diagn., Assess. Dis. Monit. 1 (2), 206–215.
Suk, H.I., Lee, S.W., Shen, D., Alzheimer's Disease Neuroimaging Initiative, 2014. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage 101, 569–582.
Suykens, J.A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9 (3), 293–300.
Tran, L., Liu, X., Zhou, J., & Jin, R. (2017). Missing modalities imputation via cascaded residual autoencoder. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (pp. 1405–1414).
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525.
Xiang, L., Li, Y., Lin, W., Wang, Q., Shen, D., 2018. Unpaired deep cross-modality synthesis with fast training. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer,, Cham, pp. 155–164.
Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., & Ye, J. (2013, August). Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 185–193).
Zhang, S., 2012. Nearest neighbor selection for iteratively kNN imputation. J. Syst. Softw. 85 (11), 2541–2552.
Zhang, T., Zhou, Z.H., 2019. Optimal margin distribution machine. IEEE Trans. Knowl. Data Eng. 32 (6), 1143–1156.
Zhu, X., Thung, K.H., Adeli, E., Zhang, Y., & Shen, D. (2017, September). Maximum mean discrepancy based multiple kernel learning for incomplete multimodality neuroimaging data. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 72–80). Springer, Cham.



Maryamossadat Aghili received her Ph.D. degree this Fall of 2021 from the School of Computing and Information Sciences (SCIS), Florida International University (FIU), Miami, USA. Prior to that, she received her Bachelor degree in 2007 in Computer Engineering from Amirkabir University of Technology-Tehran Polytechnic (AUT), Iran and her Master in Industrial Engineering in 2010. She has over 8 years of research and development experience in computer science and information systems. She also served as a research assistant with the Center for Advanced Technology and Education (CATE). She is currently working as an applied scientist at Microsoft and her focus is on deep learning for computer vision applications.



Solale Tabarestani earned her B.Sc. degree in Electrical Engineering and Electronics at 2007 and her Master degree in Electrical Engineering from Shahid Beheshti University in 2011. She obtained her Ph.D. degree Ph.D. degree in the Fall of 2021 with the Department of Electrical and Computer Engineering. She has served as a research assistant with the Center for Advanced Technology and Education (CATE) at Florida International University. Her research interests include machine learning, deep learning and computer vision. Her research focus was in developing machine learning algorithms for the study of Alzheimer's disease using cross-sectional and longitudinal studies in a multimodal neuroimaging platform. She is currently working as an applied scientist at Amazon.

Malek Adjouadi is the Ware Professor with the department of Electrical and Computer Engineering and is the founding director of the Center for Advanced Technology and Education (CATE) at Florida International University. He is also a Distinguished University Professor at FIU. He obtained his BS degree from Oklahoma State University (1978) and his MS and Ph.D. degree in Electrical Engineering from the University of Florida (1981 and 1985). While in Gainesville, Fl, he also worked as an Engineer in Hardware Diagnostics for Television Switching Boards with Vital Industries, Inc. While serving as an Assistant Professor with the University of Hawaii, he had the privilege to testify to the US Senate on Oversight Hearing on Veterans' Health Care in Hawaii and on technology to help persons with disabilities. His research interests are in image processing, neuroimaging, and assistive technology research to help persons with visual and motor disabilities. He also serves as consultant to Baptist Hospital, Mount Sinai Medical Center and the University of Miami.