Subject Harmonization of Digital Biomarkers: Improved Detection of Mild Cognitive Impairment from Language Markers

Bao Hoang[1,‡], Yijiang Pang[1,‡], Hiroko H. Dodge[2], Jiayu Zhou[1,†]

[1] *Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA*

[2] *Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA02129, USA*

[†] *Corresponding E-mail: jiayuz@msu.edu*

[‡] *Equal contribution*

Mild cognitive impairment (MCI) represents the early stage of dementia including Alzheimer's disease (AD) and is a crucial stage for therapeutic interventions and treatment. Early detection of MCI offers opportunities for early intervention and significantly benefits cohort enrichment for clinical trials. Imaging and *in vivo* markers in plasma and cerebrospinal fluid biomarkers have high detection performance, yet their prohibitive costs and intrusiveness demand more affordable and accessible alternatives. The recent advances in digital biomarkers, especially language markers, have shown great potential, where variables informative to MCI are derived from linguistic and/or speech and later used for predictive modeling. A major challenge in modeling language markers comes from the variability of how each person speaks. As the cohort size for language studies is usually small due to extensive data collection efforts, the variability among persons makes language markers hard to generalize to unseen subjects. In this paper, we propose a novel subject harmonization tool to address the issue of distributional differences in language markers across subjects, thus enhancing the generalization performance of machine learning models. Our empirical results show that machine learning models built on our harmonized features have improved prediction performance on unseen data. The source code and experiment scripts are available at https://github.com/illidanlab/subject_harmonization.

Keywords: Mild Cognitive Impairment; Harmonization Algorithm

1. Introduction

Alzheimer's disease (AD) is a major type of dementia and ranks as the seventh-leading cause of death in the United States in 2020.[1] Mild Cognitive Impairment (MCI) is the prodromal stage of dementia, including AD, characterized by minor problems with memory, language, or judgment. Early detection of MCI is critical for early intervention and cohort enrichment. *In vivo* biomarkers such as Aβ-amyloid identified by cerebrospinal fluid Aβ42 or PET amyloid imaging are sensitive to the early or pre-clinical stage. Yet, it is not easily accessible nor affordable for massive screening of general older adults, especially those with limited healthcare access.

Recently developed digital biomarkers have offered an affordable and non-intrusive alter-

native. Especially language markers,[2–4] linguistic and speech variables derived from conversations, both structured[5] or semi-structured,[4] have shown a significant correlation with the cognitive capability of the subjects and are recently used for MCI detection.[6] Digital biomarkers are generally derived and utilized in a data-driven fashion. For example, language markers are derived from carefully designed cohorts[4,7] to build predictive models that take language features as input and clinical variables as output.

One significant challenge of digital biomarkers is the limited cohort sample size, where specially designed collection protocols and devices must be deployed for data collection. For example, in the studies of language markers, the I-CONECT study[4] collected semi-structured conversation data from 74 subjects in a five-year clinical trial, and the ADReSS data from DementiaBank has spontaneous speech of 158 subjects.[7] As the small sample size greatly limits the machine learning models that can be used for analysis, a standard to enrich the sample size is constructing multiple data points from the same subject and associated with the same clinical label of the subject as the prediction target. In sensor studies, for example, by using a fixed time window, multiple time series are derived from the same subject as data points.[8,9] Another example is in language marker studies, where linguistic and speech markers are derived from one conversation, and thus multiple conversations from the same subject are treated as different data points.[2–4]

Even though these treatments greatly increased the sample size for predictive modeling, they have violated the basic assumption of most analytic approaches, that data points should be independent and identically distributed (i.i.d.). The non-*i.i.d.* is complicated by another challenge of digital biomarkers, which usually have high individual variability compared to other biomarkers, leading to unstable prediction performance and poor generalization performance to unseen subjects.[10] Again use language markers as an example: the way people speak can be drastically different, and such differences are much more outstanding than subtle differences characterizing cognitive capabilities. The intuitive idea is to harmonize the distributional bias from subjects, similar to the harmonization that removes confounding factors from demographic data or eliminates batch effects. However, *subject harmonization* has drastically distinguished itself from eliminating typical confounding variables: the subjects in the testing/inference stage are not accessible during the training, and the embedding of subject information is implicit and may be non-linearly correlated with multiple dimensions in the original feature representations. Therefore, the existing harmonization approach cannot be used to quantify and remove the subject effects.

In this paper, we propose a novel framework for subject harmonization. The proposed approach uses an auxiliary classification task on the subjects to learn a deep harmonization network, which eliminates both linear and non-linear effects in differentiating subjects. Our empirical results show that the language markers harmonized by the proposed approach can improve MCI detection performance.

2. Related Works

Detection of MCI. There are many approaches developed for detecting MCI using a combination of clinical information,[11] brain imaging,[12–15] and genetics.[16,17] For example, machine

learning models built on brain imaging such as MRI and FDG PET have been shown effective for capturing structural and metabolism information of the brain and are strongly associated with the development of AD.[14,18] Yet these biomarkers are often expensive and instructive, making them hard to screen general older adults. More recently, digital biomarkers[2–4] have offered a promising affordable, and non-intrusive alternative for broader adoption. The development of language markers is still in its early stage. Digital markers derived from the behavior are highly variable and different language markers derived from limited data often yield unstable detection models and are hard to generalize to unseen populations.

Data Harmonization. A fundamental challenge of data analysis is the harmonization of confounding variables, i.e., eliminating the effects from confounding variables.[19,20] With explicit confounding variables, common harmonization approaches eliminate confounding variables' influence on the original input features or output.[21,22] Recent deep learning models require the harmonization of non-linear effects, leading to the development of end-to-end frameworks that cooperate with the task prediction loss and a penalty loss that usually minimizes dependence between confounders and prediction outcomes.[23–26] Meanwhile, fair machine learning schemes exploit distributional robust optimization to control implicit demographic confounding effects (bias).[27–29] From another aspect, the underlying variables can be considered as some strong signal in the original features but is irrelevant to our prediction goal, then feature engineering helps reduce the effects.[30] Most existing harmonization approaches need confounding variables to be accessible during the training and secure the generalization to unseen groups. However, in digital biomarker studies where subjects are treated as a confounding variable, the challenging arises when testing subjects are not seen during the training and demands a generalizable harmonization on subjects.

3. Methods

3.1. *Data*

We use semi-structured conversational data from a clinical trial I-CONECT (Clinicaltrials.gov: NCT02871921). The data is available upon request at https://www.i-conect.org/. This clinical trial aims to investigate the potential benefits of regular video chat conversations on the cognitive functions and psychological well-being of individuals aged 75 and older. The dataset has 6771 conversation sessions from 74 participants, with 36 participants being cognitively normal (NL) and 38 diagnosed with mild cognitive impairment (MCI). Each conversational session is about 30 minutes in length. Table 1 shows the participants' demographic information.

Table 1. Demographics of Participants

Variable	All (n = 74)	NL (n = 36)	MCI (n = 38)
Age	80.7 ± 4.6	79.7 ± 3.9	81.7 ± 5.0
Gender (%women)	71.6	77.8	65.8
Years of education	15.2 ± 2.5	15.4 ± 2.5	15.1 ± 2.5
Number of Conversations	91.5 ± 37.2	92.4 ± 35.8	90.7 ± 38.4

3.2. *Language Markers*

We derived a total of 99 feature variables for each conversation as language markers, including four types: Linguistic Inquiry and Word Count (LIWC), Syntactic Complexity, Lexical Diversity, and Response Length.

Linguistic Inquiry and Word Count (LIWC): For the LIWC feature variables, we use the 2007 English version of Linguistic Inquiry and Word Count.[31] This tool categorizes English words into 64 different "LIWC categories". These categories cover a wide range of linguistic, psychological, and topical aspects, enabling us to gain insights into various social, cognitive, and affective processes. To obtain the LIWC features, we follow:[3] We first generate a 64-dimensional LIWC feature vector for every word in each conversation, with each dimension corresponding to a specific LIWC category (1 = word belongs to the category, 0 = word does not belong); we then sum over the feature vectors of all words in the conversation, resulting in a single 64-dimensional feature vector representing the linguistic feature of that conversation.

Syntactic complexity represents the range and intricacy of grammatical structures employed in language production.[32] We used the L2 Syntactic Complexity Analyzer[33] to extract the syntactic complexity feature. This tool is specifically designed to automate the analysis of syntactic complexity in English language texts produced by advanced learners of English. We extract a 23-dimensional vector from each conversation representing the syntactic complexity of conversation, with each dimension corresponding to a specific English syntactic complexity measure from the tool.

Lexical Diversity is the range of different words within a given text, wherein a wider range indicates greater diversity.[34] Given a text input, lexical diversity has been measured using the type-token ratio (TTR),[35] obtained by dividing the total number of unique words by the overall word count. To adopt this in our study, we extract the TTR from participants' conversational responses, as well as its variations, such as the moving average type-token ratio (MATTR)[36] and the mean segmental type-token ratio (MSTTR). We also use additional lexical diversity measures, including the Hypergeometric distribution D (HD-D) and the measure of textual Lexical Diversity (MTLD).[37] In total, we derive a 10-dimensional vector representing conversations' lexical diversity, with each dimension corresponding to one of the aforementioned lexical diversity measures and its respective variation.

Response length: Our analysis suggests that NL individuals tend to provide lengthier responses to questions posed by interviewer than MCI individuals, showing great potential for distinguishing between MCI and NL individuals. We extract the mean and variance of participants' response lengths within each conversation.

3.3. *Generalized Least Squares*

Generalized least squares is a widely used harmonization approach to remove linear effects given confounding variables, such as age, gender, and education.[21,38] For each conversation's extracted language marker features x_i, we assume that these features are linearly biased by three confounding variables age, sex, and education of the subject, denoted

$c_i = [\text{age}, \text{sex}, \text{education}]$, such that:

$$x_i = w \cdot c_i^T + x_i^{\text{harmonized}}$$

where w is weight matrix and $x_i^{\text{harmonized}}$ is our goal harmonized language markers. The objective function for generalized least square method is given by:

$$\min_w \sum_{i=1}^{n} (wc_i^T - x_i)^2$$

After obtaining weight matrix w by solving the above objective function, the harmonized language markers is derived by:

$$x_i^{\text{harmonized}} = x_i - w \cdot c_i^T$$

3.4. *Subject Harmonization for Non-linear Predictive Modeling*

Unlike other types of *in-vivo* biomarkers, digital markers show great individual variability. In language markers, for instance, how one speaks a language can differ greatly, even if they are all native speakers. The differences can be visualized by checking the distributions of language features. Our empirical results in Sec. 4.1 show that the feature variables have clear clustering structures w.r.t. subjects. As such, successful analysis and predictive modeling need careful harmonization to eliminate individual variability. Generalized least squares's harmonization mechanism eliminates the linear subspace that is predictive of these confounding variables and uses the orthogonal complement subspace as the harmonized features. Though all linear effects are removed through the harmonization approach, the approach does not remove any non-linear effects from data. For example, if the multiplication of two confounding variables (e.g., age and gender) has effects on the data, such effects will not be removed and will be picked up by non-linear models such as random forest and deep learning models. Another challenge comes from the generalization of harmonization, where digital biomarkers demand a unique harmonization procedure that can be generalized to unseen subjects.

To address the above challenge, we propose a deep harmonization network to facilitate analytics with digital biomarkers. In the context of the prediction of MCI from language markers, we are given a set of conversations collected from a set of different subjects and we would like to build a predictive model for MCI using these conversations. We follow the last section to extract features for *each conversation* and form a feature vector for each conversation. The setting of predictive modeling is to classify each conversation/feature vector into a label (MCI or not), which will be later aggregated into a prediction of the subject. The feature vectors of one subject will be either used in training or testing but not both. The goal of harmonization is to remove the confounding factor of subjects in the feature vectors. The proposed approach has two stages: in the first stage, we construct an auxiliary task to learn the deep harmonization network; in the second, the learned harmonization network is used to transform the data points, and the harmonized data is then used for building a downstream classifier of MCI.

The design of a deep harmonization network is based on two intuitions: 1) a good harmonization should remove all linear and non-linear effects from subjects, and therefore the harmonized features should not be able to differentiate subjects under deep models; 2) the

harmonized features should be as close to the original feature as possible (otherwise, the harmonization admits a trivial solution where all features are wiped and set to the same value). Following these intuitions, the proposed approach seeks to minimize the subject differentiation between data points obtained from different subjects and minimizes the differences between harmonized and original language markers. Generally, for M pairs of extracted language features and corresponding subject labels $(\mathbf{x}_i, \mathbf{y}_i^s)$, we denote $f_{\text{FH}}(\cdot) : \mathbf{x} \rightarrow \bar{\mathbf{x}}$ as the feature harmonization network parameterized with θ_{FH}, $f_{\text{s}}(\cdot) : \bar{\mathbf{x}} \rightarrow \mathbf{s}$ as the auxiliary subject classifier parameterized with θ_s. The composite function $f_s \circ f_{\text{FH}}$ denotes a classifier f_s using harmonized features f_{FH}. The objective for learning feature harmonization is given by:

$$\min_{\theta_{FH}, \theta_s} \frac{1}{M} \sum\nolimits_{i=1}^{M} -\ell_{ent}(f_s \circ f_{\text{FH}}(\mathbf{x}_i), \mathbf{y}_i^s) + \ell_{mse}(f_{\text{FH}}(\mathbf{x}_i), \mathbf{x}_i), \tag{1}$$

where $\ell_{ent}(\cdot)$ is the cross-entropy loss and minimizing $-\ell_{ent}(\cdot)$ encourages the harmonized features cannot be differentiated by subject identities, and $\ell_{mse}(\cdot)$ is the mean square error which encourages the similarity between the original features and the harmonized features. Note that we do not restrict the type of classifier to be used in f_s, but a non-linear model is preferred due to the design of deep harmonization. In our study, we use a 3-layer MLP for the harmonization network.

3.5. *MCI Detection using Harmonized Features*

After the harmonization process, we use the harmonized features with confounding effects removed for the downstream task of MCI detection. The MCI detection can be modeled by two classification tasks: a) *conversation classification* that identifies whether a given conversation is from an MCI subject or an NL subject using language markers extracted from the conversation, and b) *subject classification*, which collectively uses the results from the conversation classification on conversations from one subject and predict if a subject is an MCI subject or an NL subject. We model conversation classification as a standard machine learning task that seeks a classifier that takes language markers as an input and outputs a binary prediction. Formally, we have M pairs of extracted features and corresponding cognitive status label $(\mathbf{x}_i, \mathbf{y}_i^c)$. We denote $f_{\text{t}}(\cdot) : \bar{\mathbf{x}} \rightarrow \mathbf{t}$ as the MCI classifier parameterized with θ_t. In our study, we use two classifiers: a linear model (logistic regression, LR) and a non-linear model (2-layer multi-layer perceptron, MLP). Then, the objective function for cognitive status classification is formulated as:

$$\min_{\theta_t} \frac{1}{M} \sum\nolimits_{i=1}^{M} \ell(f_t \circ f_{\text{FH}}(\mathbf{x}_i), \mathbf{y}_i^c),$$

where $\ell(\cdot)$ is the binary cross entropy loss. To achieve subject classification, we use a majority vote strategy so that if more than 50% of a subject's conversations are predicted as MCI by the conversation classifier, we classify that subject as MCI and NL otherwise. For both settings, we randomly sample 80% subjects as train subjects and the remaining subjects as test subjects. The conversations from training subjects are used to train the conversation classifier. The complete framework is illustrated in Figure 1.

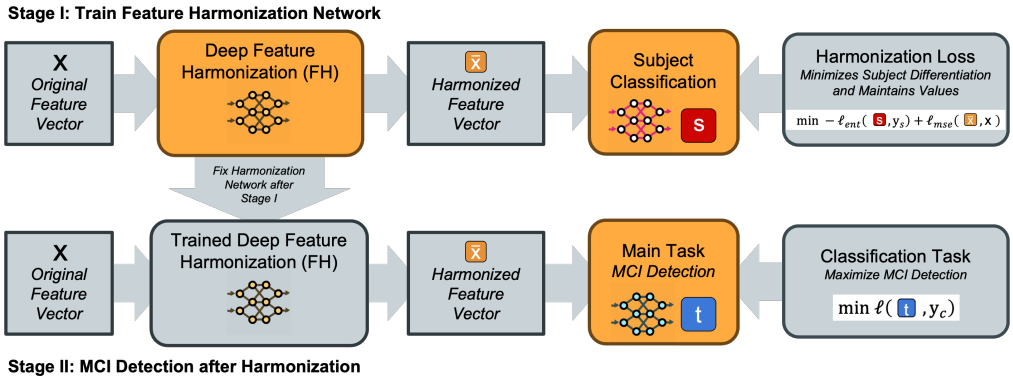

Fig. 1. The proposed subject harmonization process includes two stages. In the first stage, we train a deep harmonization network using an auxiliary subject classification task, which discourages differentiation among subjects and meanwhile retains the similarity between the original features and the harmonized ones. In the second stage, we fix the harmonization model and use the harmonized features to train the main learning task, i.e., the detection of MCI.

4. Experimental Results and Analysis

4.1. *Effectiveness and Generalizability of Subject Harmonization*

The design of harmonization is to remove the confounding factor of the variable of subjects. Therefore, we investigate the prediction power towards subjects using features before and after harmonization. The stronger the confounding variable, the better the features' prediction power differentiating subjects. A successful harmonization should greatly eliminate such prediction power.

In this experiment, conversations from individual subjects are assigned the same labels, while conversations from different subjects are assigned distinct labels. For example, all conversations from the first subject have the label 1, and all those from the second subject have the label 2. With a total of 74 subjects, we have 74 unique labels. We randomly split data (original or harmonized) into training and testing, with 80% of conversations for training and 20% for testing. We build a linear classifier (Logistic Regression) and a deep classifier (Multi-layer perceptron) using the training data and evaluate the performance in terms of accuracy using the data. For the harmonization network, we use a 3-layer Multi-layer Perceptron. We repeat the experiment for 100 random seeds, and report the average accuracy of predicting testing conversations' subject labels before and after harmonization in table 2. We use the same training and testing conversations for each random seed while evaluating before and after harmonization. We see a substantial decrease in subject classification performance in both models, showing the effectiveness of the harmonization design that removes the confounding variables' linear and non-linear effects.

We conduct a qualitative study that visualizes the distributions of the language markers before and after the subject harmonization in Figure 2. We use t-SNE[39] to plot the 99-dimensional language markers in a comprehensible 2-dimensional space, where conversations from the same subjects are assigned matching colors. From the visualization, we see that data points from the same subjects show a clear clustering structure of subjects, indicating subject

Table 2. Performance of subject classification tasks before and after subject harmonization.

Classifier	Before harmonization	After harmonization
Logistic Regression	0.921±0.007	0.221±0.012
Multi-layer Perceptron	0.905±0.007	0.219±0.038

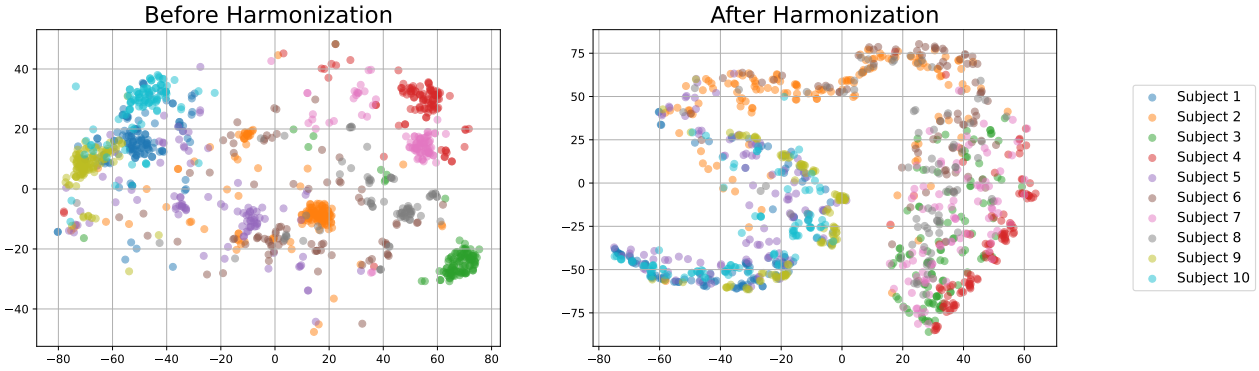

Fig. 2. The visualization of language markers extracted from conversations collected from 10 randomly selected subjects before and after subject harmonization. We see that a clear clustering structure exists before subject harmonization, which is successfully destroyed by the harmonization.

bias in the language markers. After the harmonization, such clustered structure is visually destroyed, showing the effectiveness of the purpose harmonization strategy.

4.2. *MCI Detection via Harmonized Language Markers*

We now investigate the predictive power of language markers in detecting MCI subjects. We compare a set of different harmonization approaches: a) generalized least squares,[21,38] commonly used for harmonizing linear effects and used age/gender/education as confounding variables; b) the proposed deep subject harmonization, which harmonizes against the subject variable but does not use demographic variables (age/gender/education); c) deep harmonization that does not use subject information and jointly harmonizes all demographic variables. d) deep harmonization approaches that harmonize only individual demographic variables.

When harmonizing demographic variables using a deep harmonization network, we construct category variables from age/gender/education (e.g., age between 75-79 as category 1, age between 80-84 as category 2) and train equation 1. We repeat the experiments for 100 random seeds and report the average and standard deviation of Area under the ROC curve (AUC), F1, Sensitivity, and Specificity on the test data in Table 3.

From the results, we find the following: 1) The non-linear model MLP using features from deep subject harmonization, which harmonizes the subject variable using a deep model, provides the best downstream classification performance on both conversation and subject predictions. 2) Both the linear and non-linear models benefit more from deep subject harmonization than generalized least squares. 3) For MLP, deep harmonization on demographic

Table 3. Performance of two cognitive status classification tasks over different harmonization methods.

Method for harmonization	Task Classifier	Performance metrics			
		AUC	F1	Sensitivity	Specificity
Conversation classification					
None	LR	0.583±0.098	0.557±0.092	0.570±0.123	0.557±0.101
	MLP	0.594±0.092	0.556±0.088	0.545±0.116	0.611±0.091
Generalized least squares[21]	LR	0.567±0.110	0.537±0.104	0.538±0.134	0.570±0.119
	MLP	0.545±0.109	0.522±0.103	0.516±0.132	0.574±0.125
Deep harmonization - subject (Proposed method)	LR	0.640±0.097	0.581±0.089	0.575±0.129	0.625±0.132
	MLP	0.646±0.092	0.558±0.101	0.541±0.136	0.640±0.126
Deep harmonization (- age & gender & education year)	MLP	0.527±0.120	0.517±0.119	0.593±0.227	0.427±0.235
Deep harmonization - age	MLP	0.596±0.107	0.538±0.101	0.535±0.166	0.608±0.178
Deep harmonization - gender	MLP	0.554±0.110	0.551±0.110	0.635±0.209	0.426±0.208
Deep harmonization - education year	MLP	0.611±0.102	0.589±0.080	0.654±0.141	0.477±0.165
Subject classification					
None	LR	0.591±0.124	0.579±0.126	0.593±0.166	0.568±0.169
	MLP	0.626±0.122	0.593±0.124	0.576±0.153	0.649±0.159
Generalized least squares[21]	LR	0.585±0.129	0.529±0.148	0.519±0.187	0.601±0.164
	MLP	0.568±0.122	0.568±0.138	0.565±0.175	0.605±0.175
Deep harmonization - subject (Proposed method)	LR	0.649±0.121	0.592±0.115	0.575±0.157	0.652±0.162
	MLP	0.657±0.113	0.571±0.118	0.546±0.152	0.655±0.152
Deep harmonization (- age & gender & education year)	MLP	0.538±0.148	0.539±0.165	0.637±0.272	0.381±0.282
Deep harmonization - age	MLP	0.614±0.122	0.577±0.133	0.585±0.205	0.603±0.217
Deep harmonization - gender	MLP	0.571±0.128	0.579±0.139	0.676±0.230	0.409±0.244
Deep harmonization - education year	MLP	0.639±0.122	0.632±0.091	0.736±0.159	0.417±0.218

Abbreviations: LR, Logistic Regression; MLP, Multi-layer Perceptron.

variables performs worse than generalized least squares, even though both jointly harmonize against all three demographic variables.

4.3. *Performance on Different Sub-Populations*

Table 4 presents the performance of conversation and subject classification on different sub-populations, i.e., different gender groups, education levels, and age groups. By zooming in on the performance of different sub-population groups, we want to inspect how the proposed subject harmonization impacts these groups, given that demographic variables are not used in the harmonization process. From the results, we see that the proposed subject harmonization consistently improved the performance of most groups, with the exception of 1) the higher educated group (Edu years 19-21), for both conversation and subject classification, and 2) minor performance drop in the Male group for the subject classification.

Table 4. Performance of two cognitive status classification tasks before and after the harmonization methods.

Groups	Performance compairsion					
	Before harmonization			After harmonization		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Conversation classification						
Male	0.533±0.185	0.517±0.199	0.537±0.213	0.564±0.199	0.656±0.228	0.409±0.257
Female	0.621±0.112	0.554±0.140	0.641±0.103	0.673±0.106	0.475±0.181	0.728±0.143
Edu 12-15	0.483±0.162	0.529±0.182	0.447±0.165	0.618±0.186	0.586±0.205	0.599±0.234
Edu 16-18	0.621±0.163	0.490±0.185	0.715±0.110	0.668±0.146	0.452±0.241	0.735±0.160
Edu 19-21	0.857±0.096	0.790±0.182	0.743±0.161	0.732±0.323	0.647±0.418	0.498±0.257
Age 75-80	0.608±0.123	0.532±0.158	0.648±0.096	0.638±0.111	0.519±0.169	0.664±0.130
Age 81-87	0.500±0.231	0.483±0.224	0.529±0.317	0.517±0.309	0.456±0.263	0.512±0.369
Age 88-94	0.781±0.189	0.918±0.157	0.339±0.129	0.941±0.058	0.987±0.026	0.386±0.293
Subject classification						
Male	0.589±0.275	0.537±0.299	0.587±0.365	0.577±0.277	0.641±0.292	0.384±0.392
Female	0.653±0.152	0.600±0.184	0.665±0.187	0.691±0.158	0.491±0.204	0.751±0.184
Edu 12-15	0.480±0.211	0.530±0.226	0.377±0.291	0.624±0.215	0.601±0.218	0.603±0.247
Edu 16-18	0.694±0.241	0.549±0.327	0.828±0.199	0.699±0.228	0.445±0.295	0.756±0.221
Edu 19-21	1.000±0.000	0.929±0.258	0.921±0.260	0.754±0.395	0.607±0.457	0.508±0.445
Age 75-80	0.654±0.176	0.561±0.206	0.715±0.185	0.671±0.153	0.512±0.212	0.699±0.158
Age 81-87	0.515±0.309	0.501±0.331	0.569±0.431	0.541±0.379	0.474±0.320	0.536±0.444
Age 88-94	0.953±0.192	0.984±0.087	0.141±0.336	0.984±0.087	1.000±0.000	0.328±0.426

4.4. *Important Language Markers Before and After Harmonization*

In this section, we investigate the feature importance and compare the top language markers before and after harmonization. For linear models, feature importance can be directly derived from the model weights, and for non-linear MLP models used in this paper, we do not have such a straightforward way of getting them. We adopt commonly used permutation feature importance[40] to estimate the feature importance. We permute each feature's values and subsequently feed the modified dataset into our pipeline. After that, we derive the AUC score for both conversation and subject classification using this permutated dataset. The feature importance of a feature is then determined by computing the difference between the AUC values obtained from the original dataset and the permuted dataset. A larger decrease in AUC indicates higher importance of the respective feature in the classification model.

In table 5, we present the top 10 language features before and after the feature harmonization for both conversation and subject classification. We see that: 1) top features differ quite much before and after harmonization. Notably, we see "Nonfluencies" being the most important feature after harmonization, which better supports the pathology of dementia, where dementia (even at the preclinical stage) may impact a subject, making it harder to find the right words and therefore showing a higher number of nonfluencies during communication. 2)

Table 5. Top 10 language features before and after harmonization where the importance is w.r.t. the decreasing of AUC in both conversation classification and subject classification.

Before harmonization			After harmonization		
Feature name	Type	AUC drop	Feature name	Type	AUC drop
Conversation classification					
Negations	LIWC	0.02749	Nonfluencies	LIWC	0.00616
1st pers plural	LIWC	0.00587	Assent	LIWC	0.00471
Discrepancy	LIWC	0.00495	Insight	LIWC	0.00468
Assent	LIWC	0.00328	Affective processes	LIWC	0.00455
Family	LIWC	0.00325	T-unit per sentence	SC	0.00455
Tentative	LIWC	0.00324	3rd pers singular	LIWC	0.00439
Sexual	LIWC	0.00297	Causation	LIWC	0.00435
Auxiliary verbs	LIWC	0.00238	Certainty	LIWC	0.00418
Home	LIWC	0.00215	Mean length of sentence	SC	0.00414
Inhibition	LIWC	0.00204	Hear	LIWC	0.00395
Subject classification					
Negations	LIWC	0.03469	T-unit per sentence	SC	0.01203
Tentative	LIWC	0.00562	Mean length of sentence	SC	0.00969
Family	LIWC	0.00547	Negations	LIWC	0.00922
Textual lexical diversity	LD	0.00531	Clause	SC	0.00906
Home	LIWC	0.00438	Affective processes	LIWC	0.00859
Social processes	LIWC	0.00391	Causation	LIWC	0.00844
1st pers plural	LIWC	0.00359	Cognitive processes	LIWC	0.00828
Assent	LIWC	0.00344	Positive emotion	LIWC	0.00813
Personal pronouns	LIWC	0.00313	Inclusive	LIWC	0.00813
Discrepancy	LIWC	0.00313	Motion	LIWC	0.00750

Abbreviations: LIWC, Linguistic Inquiry and Word Count; SC, Syntactic Complexity; LD, Lexical Diversity.

more syntactic complexity features appear after harmonization for subject classification. The top features "T-unit per sentence" and "mean length of sentence" directly correlate to the language capability of constructing longer features.

5. Discussion

In this paper, we propose a subject harmonization algorithm to mitigate the distributional difference of digital biomarkers induced by subject variability. Our empirical results show that applying subject harmonization to language markers improves the performance of MCI detection. We show the effects of subject variability from a quantitative perspective using a subject prediction task, and also from a qualitative perspective from visible clusters in the visualization of language markers. Our experiments show that the proposed subject harmonization approach effectively mitigates the subject variability so that the harmonized data has much less power to differentiate among subjects. Meanwhile, we show that MCI detection models

built from language markers harmonized by the proposed subject harmonization improve the predictive performance. The harmonization improves the AUC score of MCI prediction from 0.594 to 0.646 in conversation classification task and from 0.626 to 0.657 in subject classification task. We further investigated the sub-group performance of different age/gender/years of education, and we see that the performance of most groups have been improved.

Despite the improvement in prediction performance using language markers through the harmonization algorithm, future studies still need investigation. Firstly, the prediction performance from language markers is yet to be improved. A possible reason is the quality of the language markers and that we only used linguistic and syntactic information. We will study subject harmonization on additional feature variables, such as speech and video. Secondly, performing subject harmonization on demographic variables witnessed reduced predictive performance, indicating that the proposed deep harmonization network is currently not applicable to general harmonization usage. We plan to investigate theoretical relationship between the two harmonization types, and improve deep harmonization network to handle demographic variables. Thirdly, while we have successfully validated the positive impact of harmonization on language markers, it remains to confirm its efficacy on other data types. We plan to dedicate considerable time to applying the harmonization algorithm to different types of markers, such as clinical data or brain imaging data. This broader exploration will enable us to assess the generalizability and versatility of the harmonization technique across various data modalities, facilitating a more comprehensive understanding of its potential applications.

6. Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant IIS-2212174, IIS-1749940, Office of Naval Research N00014-20-1-2382, and National Institute on Aging (NIA) RF1AG072449, R01AG051628, R01AG056102.

References

1. S. L. Murphy, K. D. Kochanek, J. Xu and E. Arias, Mortality in the united states, 2020 (2021).
2. B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead and J. Kaye, Spoken language derived measures for detecting mild cognitive impairment, *IEEE Transactions on Audio, Speech, and Language Processing* **19**, 2081 (September 2011).
3. F. Tang, J. Chen, H. H. Dodge and J. Zhou, The joint effects of acoustic and linguistic markers for early identification of mild cognitive impairment, *Frontiers in digital health* **3**, p. 702772 (2022).
4. M. Asgari, J. Kaye and H. Dodge, Predicting mild cognitive impairment from spontaneous spoken utterances, *Alzheimer's & Dementia—Translational Research and Clinical Intervention* **3**, 219 (February 2017).
5. M. L. Manning, Improving clinical communication through structured conversation, *Nurs Econ* **24**, 268 (2006).
6. L. Chen, H. H. Dodge and M. Asgari, Topic-Based Measures of Conversation for Detecting Mild Cognitive Impairment, *Proc Conf Assoc Comput Linguist Meet* **2020**, 63 (Jul 2020).
7. S. Luz, F. Haider, S. de la Fuente, D. Fromm and B. MacWhinney, Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge, in *Proceedings of INTERSPEECH 2020*, (Shanghai, China, 2020).

8. J. Li, Y. Rong, H. Meng, Z. Lu, T. Kwok and H. Cheng, Tatc: predicting alzheimer's disease with actigraphy data (2018).
9. X. Ouyang, Design and deployment of multi-modal federated learning systems for alzheimer's disease monitoring (2023).
10. J. Yang, K. Zhou, Y. Li and Z. Liu, Generalized out-of-distribution detection: A survey (2021).
11. J. Venugopalan, L. Tong, H. R. Hassanzadeh and M. D. Wang, Multimodal deep learning models for early detection of alzheimer's disease stage, *Scientific reports* **11**, p. 3254 (2021).
12. J. Zhou, L. Yuan, J. Liu and J. Ye, A multi-task learning formulation for predicting disease progression (2011).
13. J. Zhou, J. Liu, V. A. Narayan, J. Ye, A. D. N. Initiative *et al.*, Modeling disease progression via multi-task learning, *NeuroImage* **78**, 233 (2013).
14. Q. Wang, L. Zhan, P. M. Thompson, H. H. Dodge and J. Zhou, Discriminative fusion of multiple brain networks for early mild cognitive impairment detection (2016).
15. S. Gill, P. Mouches, S. Hu, D. Rajashekar, F. P. MacMaster, E. E. Smith, N. D. Forkert and Z. I. and, Using machine learning to predict dementia from neuropsychiatric symptom and neuroimaging data, *Journal of Alzheimer's Disease* **75**, 277 (May 2020).
16. Q. Wang, M. Sun, L. Zhan, P. Thompson, S. Ji and J. Zhou, Multi-modality disease modeling via collective deep matrix factorization (2017).
17. R.-H. Lin, C.-C. Wang and C.-W. Tung, A machine learning classifier for predicting stable MCI patients using gene biomarkers, *International Journal of Environmental Research and Public Health* **19**, p. 4839 (April 2022).
18. C. Yin, S. Li, W. Zhao and J. Feng, Brain imaging of mild cognitive impairment and alzheimer's disease, *Neural regeneration research* **8**, p. 435 (2013).
19. J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou and K. Zhang, The practical implementation of artificial intelligence technologies in medicine, *Nature medicine* **25**, 30 (2019).
20. E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, *Nature medicine* **25**, 44 (2019).
21. Q. Wang, L. Guo, P. M. Thompson, C. R. Jack Jr, H. Dodge, L. Zhan, J. Zhou, A. D. N. Initiative *et al.*, The added value of diffusion-weighted mri-derived structural connectome in evaluating mild cognitive impairment: A multi-cohort validation, *Journal of Alzheimer's Disease* **64**, 149 (2018).
22. E. Adeli, X. Li, D. Kwon, Y. Zhang and K. M. Pohl, Logistic regression confined by cardinality-constrained sample and feature selection, *IEEE transactions on pattern analysis and machine intelligence* **42**, 1713 (2019).
23. Q. Zhao, E. Adeli and K. M. Pohl, Training confounder-free deep learning models for medical applications, *Nature communications* **11**, p. 6010 (2020).
24. E. Adeli, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, J. C. Niebles and K. M. Pohl, Representation learning with statistical independence to mitigate bias (2021).
25. M. Horry, S. Chakraborty, B. Pradhan, M. Paul, J. Zhu, H. W. Loh, P. D. Barua and U. R. Arharya, Debiasing pipeline improves deep learning model generalization for x-ray based lung nodule detection (2022).
26. A. Vento, Q. Zhao, R. Paul, K. M. Pohl and E. Adeli, A penalty approach for normalizing feature distributions to build confounder-free models (2022).
27. H. Namkoong and J. C. Duchi, Stochastic gradient methods for distributionally robust optimization with f-divergences, *Advances in neural information processing systems* **29** (2016).
28. T. Hashimoto, M. Srivastava, H. Namkoong and P. Liang, Fairness without demographics in repeated loss minimization (2018).
29. S. Jeong and H. Namkoong, Robust causal inference under covariate shift via worst-case subpopulation treatment effects (2020).

30. A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists* (" O'Reilly Media, Inc.", 2018).
31. The development and psychometric properties of liwc2015.
32. L. Ortega, Syntactic complexity in l2 writing: Progress and expansion, *Journal of Second Language Writing* **29**, 82 (September 2015).
33. X. Lu, Automatic analysis of syntactic complexity in second language writing, *International journal of corpus linguistics* **15**, 474 (2010).
34. M. M. Baese-Berk, S. Drake, K. Foster, D. yong Lee, C. Staggs and J. M. Wright, Lexical diversity, lexical sophistication, and predictability for speech in multiple listening conditions, *Frontiers in Psychology* **12** (June 2021).
35. W. Johnson, Studies in language behavior: A program of research, *Psychological Monographs* **56**, 1 (1944).
36. M. A. Covington and J. D. McFall, Cutting the gordian knot: The moving-average type–token ratio (mattr), *Journal of quantitative linguistics* **17**, 94 (2010).
37. P. M. McCarthy and S. Jarvis, Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment, *Behavior research methods* **42**, 381 (2010).
38. R. McNamee, Regression modelling and other methods to control confounding, *Occupational and environmental medicine* **62**, 500 (2005).
39. L. van der Maaten and G. Hinton, Visualizing data using t-sne, *Journal of Machine Learning Research* **9**, 2579 (2008).
40. L. Breiman, Random forests, *Machine learning* **45**, 5 (2001).