IISE Transactions on Healthcare Systems Engineering

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uhse21

Uncertainty-driven modality selection for data-efficient prediction of Alzheimer's disease

Zhiyang Zheng, Yi Su, Kewei Chen, David Weidman, Teresa Wu, ShihChung Lo, Fleming Lure, Jing Li & for the Alzheimer's Disease Neuroimaging Initiative

To cite this article: Zhiyang Zheng, Yi Su, Kewei Chen, David Weidman, Teresa Wu, ShihChung Lo, Fleming Lure, Jing Li & for the Alzheimer's Disease Neuroimaging Initiative (2024) Uncertainty-driven modality selection for data-efficient prediction of Alzheimer's disease, IISE Transactions on Healthcare Systems Engineering, 14:1, 18-31, DOI: 10.1080/24725579.2023.2227197

To link to this article: https://doi.org/10.1080/24725579.2023.2227197

IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
2024, VOL. 14, NO. 1, 18–31
https://doi.org/10.1080/24725579.2023.2227197

Uncertainty-driven modality selection for data-efficient prediction of Alzheimer's disease

Zhiyang Zheng[a], Yi Su[b], Kewei Chen[b], David Weidman[b], Teresa Wu[c], ShihChung Lo[d], Fleming Lure[d], Jing Li[a], and for the Alzheimer's Disease Neuroimaging Initiative[#]

[a]H. Hilton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA; [b]Banner Alzheimer's Institute, Phoenix, AZ, USA; [c]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA; [d]MS Technologies Corp, Rockville, MD, USA

ABSTRACT

Alzheimer's disease (AD) is a devastating neurodegenerative disorder. Early prediction of the risk of converting to AD for individuals at pre-dementia stages such as Mild Cognitive Impairment (MCI) is important. This could provide an opportunity for early intervention to slow down disease progression before significant irreversible neurodegeneration occurs. Neuroimaging datasets of different modalities such as MRI and PET have shown great promise. However, different data modalities are associated with varying acquisition costs/levels of accessibility to patients. We propose a machine learning (ML) framework, namely Uncertainty-driven Modality Selection (UMoS), that allows for sequentially adding data modalities for each patient on an as-needed basis, while at the same time achieving high prediction accuracy as if all the modalities were used. UMoS provides a tool to assist clinicians in deciding what data modalities/diagnostic exams each patient needs. We apply UMoS to a real-world dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) based on demographic/clinical data, MRI, and PET. UMoS shows high accuracy for predicting MCI conversion to AD, which has no significant difference from models based on the simultaneous use of all the modalities for each patient. The benefit of UMoS is significant data efficiency accomplished by saving a large percentage of patients from needing to acquire more costly/less accessible data modalities, thus lessening the burden on patients and the healthcare system.

KEYWORDS

Machine learning; multi-modality data; computer-aided diagnosis

1. Introduction

Alzheimer's disease (AD) is a devastating neurodegenerative disorder and the most common form of dementia. AD currently affects 6.2 million people aged 65 and older in the U.S. (Alzheimer's, 2022). The symptoms of AD start with mild memory loss and cognitive decline, which are followed by gradual deterioration of other brain functions. No cure for AD is currently available, but there is consensus that potential disease-modifying treatments will be more effective in slowing down cognitive decline when given at earlier stages of the disease.

Mild Cognitive Impairment (MCI) is a prodromal stage when patients show a noticeable cognitive decline, typically involving memory loss, but the symptoms are not severe enough to disrupt the ability to perform daily activities independently. The etiologies of MCI are heterogeneous, and the prognosis depends in part on the primary etiology for a given individual. Thus, some individuals with MCI will progress to AD dementia at a future time, whereas others may remain stable or even revert. It is an important task to predict if an MCI patient will progress to AD, which could provide an opportunity for early intervention to slow down disease progression before significant irreversible neurodegeneration occurs.

Neuroimaging datasets have shown great promise to predict the progression (a.k.a. conversion) of MCI to AD. Especially, neuroimages of different types/modalities measure different aspects of the brain affected by the early stage of the disease. One of the most commonly used neuroimaging modalities is T1-weighted volumetric magnetic resonance imaging (MRI), which measures brain structure. Bron et al. (2015) reviewed some earlier works using MRI for predicting MCI conversion to AD and provided benchmarking studies. In more recent years, various machine learning (ML) and deep learning methods have been adopted in this field. For example, Moradi et al. (2015) proposed a semi-supervised low-density separation model to classify MCI converters versus non-converters, with MRI features selected by regularized logistic regression. Beheshti et al. (2017) proposed a feature ranking method based on t-test scores and a genetic algorithm with Fisher criterion and used support vector machine (SVM) for classification. Zhang et al. (2021)

CONTACT Jing Li jli3175@gatech.edu H. Hilton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA.
[#]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

proposed a densely connected convolutional neural network (CNN) and applied a connection-wise attention mechanism to transform MRI into multi-level features for classification. Bron et al. (2021) proposed to preprocess MRI into a modulated grey matter map and predict MCI conversion to AD with SVM and CNN.

Compared to the research studies using MRI only, combining data from different neuroimaging modalities has demonstrated improved prediction power in studies related to AD. Positron emission tomography (PET) is another commonly used neuroimaging modality which measures various brain metabolic or biochemical processes depending on the use of radioligands. AD pathology is characterized by two pathologic hallmarks: amyloid plaques and neurofibrillary tangles (Holtzman et al., 2011). Amyloid-PET is a type of PET imaging that can show amyloid plaque deposition in the brain in the preclinical stage of AD, several years before cognitive symptoms appear. Amyloid-PET holds promise for predicting MCI conversion to AD, especially when combined with structural MRI data to exploit the complementary strength (Rosenberg et al., 2013). Some multi-modality ML methods have been proposed to integrate amyloid-PET and MRI. For example, Xu et al. (2016) proposed a weighted multi-modality sparse representation method, which minimized the weighted sum of mean squared errors of the predictions of MCI conversion by different modalities. Zhu et al. (2019) proposed a self-paced multi-kernel learning method, in which a multi-kernel linear regression with low-rank constraints on the regression coefficients was used to fuze heterogeneous modalities for classification.

To build a multi-modality ML model, it may be difficult to obtain many samples with all the modalities available. To leverage samples with partial or incomplete modalities, some ML models have been developed in AD studies. For example, Yuan et al. (2012) proposed two learning frameworks: the first framework divided data into multiple tasks based on available modalities and selected common features using sparse learning regularization; the second framework trained base classifiers on each modality to create a score matrix, imputed missing values, and integrated the modalities with a new classifier. Xiang et al. (2014) proposed a bi-level learning model which learned individual models for each modality and then integrated all the models *via* regularizations/constraints. Zhou et al. (2019) proposed a stage-wise model with each stage learning feature representations for different combinations of modalities using the maximum number of available samples. Zhou et al. (2020) proposed a modality-specific projection loss to learn a common latent space with missing modalities bypassed and build classifiers based on the latent features. Wang et al. (2020) proposed to train models on each modality independently using all the available data, which were used as teachers to help the training of the model using complete modalities. Liu et al. (2021) proposed an incomplete-multimodality transfer learning model, which built predictive models for different combinations of modalities and coupled the model estimation processes to enable transfer learning.

Our work in this study was driven by a practical consideration that different data modalities are associated with varying acquisition costs/levels of accessibility to patients. For example, among imaging modalities, PET is more costly, involves radiation exposure, and is less accessible than MRI. Also, comparing imaging and non-imaging data such as basic clinical assessments, the former may be more costly. Thus, the goal of this study is to develop a need-based approach within the context of predicting MCI conversion to AD, which sequentially adds data modalities for each patient, starting from standard ones (with lower costs and higher accessibility to patients) and gradually adding more sophisticated ones (with higher costs and lower accessibility). A modality is acquired for a patient when the predictive model based on previously acquired modalities lacks certainty. By doing so, the process for predicting MCI conversion to AD can be more "personalized" with data modalities added on an as-needed basis for each individual, instead of enforcing one-process-fits-all. In this paper, we aim to develop an ML framework, namely the Uncertainty-driven Modality Selection (UMoS) framework, to automate such a personalized, need-based process for predicting MCI conversion to AD. While there is an abundance of existing research that develops multi-modality ML models for AD, past studies mainly focused on designing different ways to integrate multi-modality datasets. Also, UMoS is different from the incomplete multi-modality learning methods that were previously reviewed. While the latter methods assume that the available modalities of each patient are fixed and given, UMoS aims to determine the minimally needed modalities for each patient in a sequential manner, which can achieve the same level of prediction accuracy as if all the modalities were used. Overall, there is no study to our best knowledge that has the same goal as UMoS, which is to achieve high predictive accuracy with patient-specific need-based data efficiency.

There are two building blocks of the UMoS framework: ML models based on sequentially added data modalities and the capacity of quantifying uncertainties of the model-based predictions. The latter drives the decision as to if the next modality should be acquired for each individual patient. Uncertainty quantification of ML models has drawn great attention recently, as it is closely related to AI safety (Abdar et al., 2021). With more and more automated systems driven by ML/AI algorithms being deployed in real-life settings, it is important to know when the algorithm is uncertain about its prediction and inform humans in order to avoid catastrophic consequences. There are two types of uncertainties that impact the prediction of ML models (Hüllermeier & Waegeman, 2021). Epistemic uncertainty refers to the uncertainty of the model structure or parameters, usually arising due to insufficient training data. Aleatoric uncertainty arises due to the measurement noise in the data itself. Both uncertainties induce predictive uncertainty, the confidence for the prediction made by an ML model. Existing research focuses on developing uncertainty quantification methods for ML (Abdar et al., 2021), uncertainty-mitigating ML models (Hariri et al., 2019), and active learning strategies (Zhang &

Lee, 2019). Different from the existing research, UMoS aims to use the uncertainty of ML prediction to drive the selection of data modalities for each individual with MCI, with the ultimate goal of predicting the individual's risk of conversion to AD with both high accuracy and high data efficiency. The contributions of this paper are summarized as follows:

- We propose an ML framework, UMoS, that allows for sequentially adding data modalities for each patient on an as-needed basis. Compared to the existing ML research in multi-modality integration, our study is unique in its goal of achieving high predictive accuracy with patient-specific need-based data efficiency.
- We propose a new model formulation in UMoS that allows the model training with incomplete modalities to distill knowledge from that with complete modalities. We further propose a strategy of using the predictive uncertainty of a model to drive the selection of data modalities for each patient. Compared to the existing research that investigates the uncertainty of ML models, UMoS is unique in the aspect of using the uncertainty to drive patient-specific modality selection, which has not been explored before to our best knowledge.
- We perform theoretical studies to show that the design of the models in UMoS respects a "more certain more accurate (MCMA)" condition. Different from typical supervised learning models that aim to maximize accuracy, the models included in UMoS should be trained such that if the model is more certain about a prediction, the prediction result should be more accurate, namely, MCMA. This is important because the predictive uncertainty (or certainty) of the model for a given patient will trigger the decision as to whether the next data modality needs to be acquired.
- We apply UMoS to a real-world dataset collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) to predict the progression to AD of MCI patients. In the two-modality case where MRI and demographic/clinical data are considered as one modality and amyloid-PET as the other modality, we demonstrate that 77% of patients can be saved from needing to acquire PET, whereas the prediction accuracy has no significant difference from the ML model based on all modalities. We also demonstrate UMoS in a three-modality case. These results show the high accuracy and data efficiency achieved by UMoS.

2. Method

2.1. Data description

This study uses the data collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) project. ADNI (http://adni.loni.ucla.edu) was launched in 2003 by the NIH, FDA, private pharmaceutical companies, and nonprofit organizations, as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. After the initial ADNI project ended, subsequent efforts known as ADNI-GO, ADNI-2 and ADNI-3 added additional participants to augment the cohort. The primary goal of ADNI has been to test whether MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. ADNI is the result of the efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. For up-to-date information, see http://www.adni-info.org/.

Our dataset was composed of 1319 samples from MCI patients in the ADNI databases. Three data modalities were included: demographic/clinical data, T1-weighted volumetric MRI, and 18 F-Florbetapir amyloid-PET (referred to as amyloid-PET in short hereafter). For the demographic/clinical dataset, we included basic demographic information such as age, gender, and education level; commonly used cognitive test scores such as the Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating Scale (CDR); status of the $\varepsilon 4$ allele of apolipoprotein E (APOE) which is a major genetic risk factor of AD. All samples have demographic/clinical data and MRI available (480 converters and 839 non-converters according to a three-year conversion time window). Only a subset of 612 samples additionally has amyloid-PET (156 converters and 456 non-converters), whereas the remaining 707 samples do not have amyloid-PET. Finally, it is worth mentioning that the 1319 samples were from 536 MCI patients since each patient may have multiple visits for data collection. Among these patients, 161 do not have amyloid-PET. We included all the samples in this study to increase the sample size. On the other hand, we were aware of the potential risk of overfitting by using this strategy. To prevent overfitting, we performed training/-validation split by patients not by samples. That is, if a patient is selected into the training (or validation) set, all the samples associated with the patient will go into the training (or validation) set. In this way, we avoided including samples from the same patient in both training and validation in order to prevent the risk of overfitting. Similar strategies have been adopted by other researchers (Zhou et al., 2019).

2.2. Image preprocessing and feature extraction

The MRI included in this study was processed by FreeSurfer v7.1 to obtain volumetric and cortical thickness measures following standard procedures (Desikan et al., 2006; Fischl et al., 2002; Fischl et al., 2004). Amyloid-PET was processed by a PET Unified Pipeline to obtain regional standardized uptake value ratios (SUVR) measurements for FreeSurfer defined regions (Su et al., 2013; Su et al., 2015). In this study, we included volumetric and thickness measures for 68 cortical regions of interest (ROIs) and volumetric measures for 14 sub-cortical and 6 ventricle regions, amounting to a total of 156 features from MRI. Also, we included quantitative measures of amyloid SUVRs with cerebella reference region for 68 cortical regions, 14 sub-cortical regions, and 68 white matter regions, and a mean SUVR feature, amounting to 151 features from amyloid-PET.

2.3. Proposed UMoS framework

2.3.1. Overview of the UMoS framework

Suppose there are M data modalities denoted by $X^{(1)}$, $X^{(2)}, \ldots, X^{(M)}$. For example, we can consider demographic/-clinical data, MRI, and PET as three modalities. $X^{(1)}$ contains demographic and clinical variables; $X^{(2)}$ and $X^{(3)}$ contain the features extracted from MRI and PET, respectively. Let y denote the class label, e.g., MCI converters or non-converters to AD. Furthermore, assume that the M data modalities are ordered such that a modality later in the order is more costly or difficult to acquire, but adding that modality will improve or at least retain the accuracy of predicting the class label for a patient compared to using all the previous modalities. To obtain such an order in a particular application, domain knowledge can be leveraged. For example, according to domain knowledge, one can reasonably order demographic/clinical data, MRI, and PET in our application as $X^{(1)}$, $X^{(2)}$, $X^{(3)}$.

If not considering the cost of data acquisition, it would be best to collect all the data modalities for every patient. However, the cost aspect cannot be overlooked in practice. Therefore, the basic idea of UMoS is to only add a modality for a patient if it is needed, where the necessity is determined using the prediction uncertainty based on all previous modalities, i.e., when this prediction uncertainty exceeds a threshold. This method imitates the diagnostic process of physicians, who would order a diagnostic test (usually more expensive, complicated, or invasive, but also more accurate) when all previous tests cannot support a decision with certainty.

Formally, the UMoS framework includes a collection of models, $f^{(1)}, \ldots, f^{(1:M)}$, that can be deployed sequentially for a given patient. Each model in this collection is in the form of $f^{(1:m)}: X^{(1:m)} \rightarrow y$, i.e., it takes collective modalities up to the m-th modality as input to predict the class label y, $m = 1, \ldots, M$. Additionally, the UMoS framework requires that we can have a score to quantify the uncertainty of each model $f^{(1:m)}$ regarding its prediction, denoted by $u^{(1:m)}$. Having both model-based prediction and uncertainty quantification imitates the decision-making process by physicians, which typically includes not only a diagnostic/prognostic result (e.g., having or not having a certain disease, will convert or will not convert to AD dementia for an MCI patient) but also the certainty/uncertainty level of the physician regarding the result. Figure 1 depicts the workflow of how the models in UMoS are sequentially deployed for a given patient. Specifically, the patient will first acquire the data modality $X^{(1)}$. Using $X^{(1)}$ as input, the first model $f^{(1)}$ will be deployed. If the uncertainty score associated with the model $f^{(1)}$ is lower than a threshold, i.e., $u^{(1)} \leq u^{*(1)}$, then the class label of the patient will be predicted using $f^{(1)}$, i.e., $\hat{y}^{(1)}$, and the workflow for this patient is completed. Otherwise, if $u^{(1)} > u^{*(1)}$, the patient will need to acquire the next data modality $X^{(2)}$. Using collective modalities $X^{(1:2)}$ as input, the second model $f^{(1:2)}$ will be deployed, and a similar step as previously described will be followed. In this way, the sequence of models, $f^{(1)}, \ldots, f^{(1:M)}$, will be deployed one after another until the first time when the uncertainty score generated by a model is less than the threshold or all the data modalities have been acquired.

2.3.2. Construction of the UMoS framework

The key to constructing the UMoS framework is to train the models $f^{(1)}, \ldots, f^{(1:M)}$. Let D_{tr} denote a training set of n samples. We do not require all the training samples to have all the modalities available, as such a dataset would be too expensive to collect in practice. Instead, let $\left(X_i^{(1:m)}, y_i\right)_{i=1,\ldots n^{(1:m)}}$ denote the samples with collective modalities $X^{(1:m)}$ available, $m = 1, \ldots M$. The sample size $n^{(1:m)}$ typically decreases as m increases.

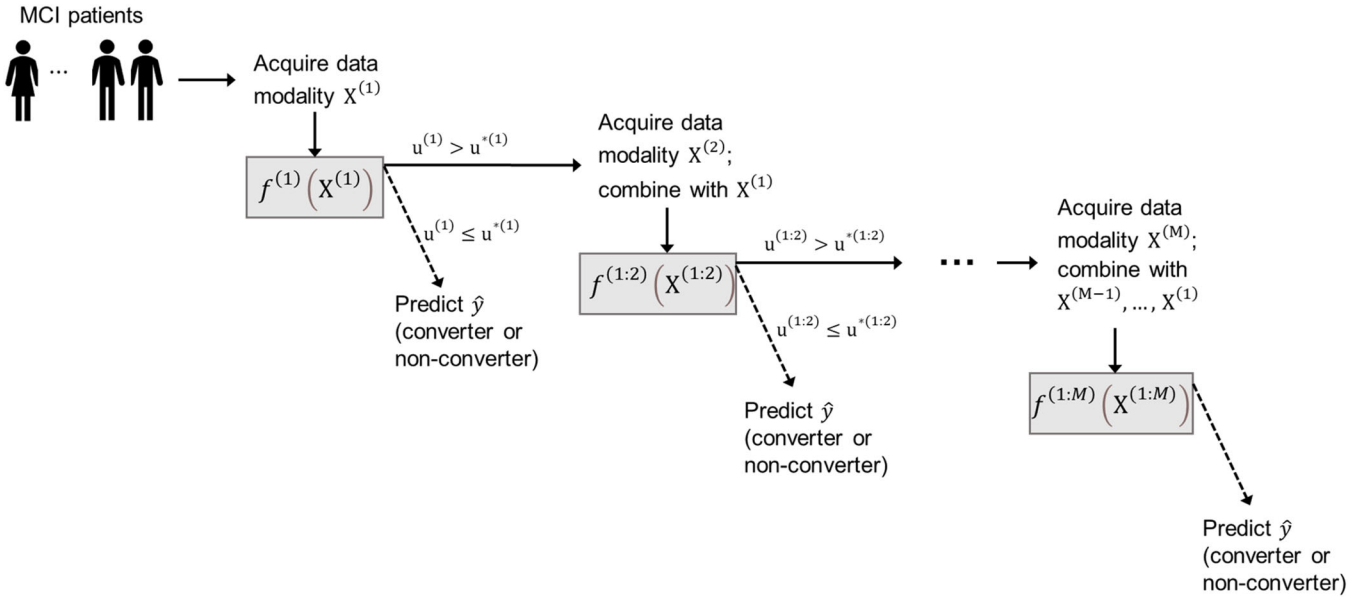

Figure 1. Overview of the UMoS-based workflow in predicting the conversion risk to AD for each MCI patient.

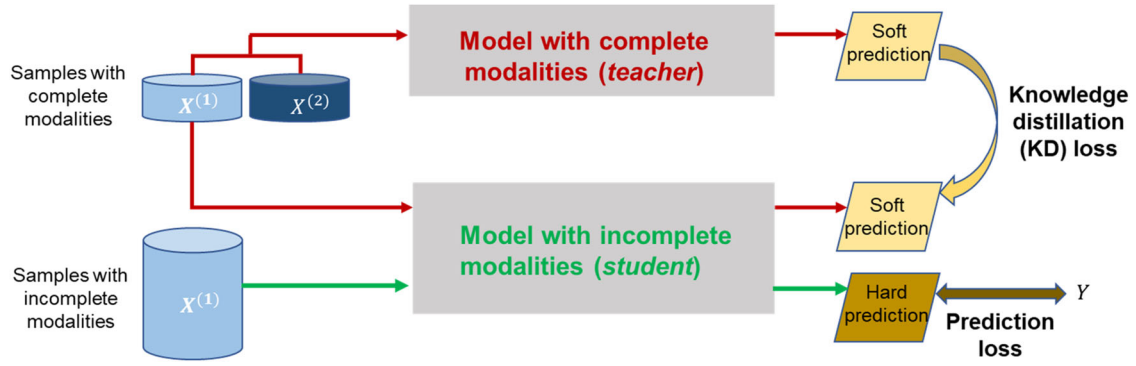

Figure 2. Graphical illustration of the incomplete-modality model in Equation (1) for a simple case of two modalities.

The models included in the UMoS framework can be divided into two categories: $f^{(1:M)}$ is the model based on complete/all modalities; $f^{(1)},\ldots,\ f^{(1:M-1)}$ are models based on incomplete (subsets of) modalities. Because of this difference, the two categories of models will be constructed in different ways. Specifically, the construction of $f^{(1:M)}$ is straightforward. In theory, it can be trained using any classification algorithm based on samples in the training set with complete modalities. In practice, one can try a variety of different algorithms and choose the one with the best accuracy to be $f^{(1:M)}$.

Building the models with incomplete modalities, i.e., $f^{(1)},\ldots,\ f^{(1:M-1)}$, needs some special considerations, which is the focus of discussion in this section. Specifically, to train each model with incomplete modalities, $f^{(1:m)}$, $m=1,\ldots M-1$, we propose the following optimization:

$$\min_{f^{(1:m)}} \overbrace{\sum_{i=1}^{n^{(1:m)}} l\left(y_i, f^{(1:m)}\left(X_i^{(1:m)}\right)\right)}^{prediction\ \ loss} + \lambda_1 \overbrace{\sum_{j=1}^{n^{(1:M)}} d\left(\tilde{f}^{(1:m)}\left(X_j^{(1:M)}\right), \tilde{f}^{(1:M)}\left(X_i^{(1:M)}\right)\right)}^{knowlege\ \ distillation\ \ (KD)} + \lambda_2 \overbrace{\|f^{(1:m)}\|}^{model\ \ complexity}, \quad (1)$$
$$m=1,\ldots M-1.$$

The first term is the prediction loss computed based on training samples with collective modalities $X^{(1:m)}$ available. The second term is to encourage the prediction by $f^{(1:m)}$ to be similar to that by the model using complete modalities, $f^{(1:M)}$, which is supposed to have the best performance. Here we used $\tilde{f}$ to denote the "soft" prediction made by a model, which is known to improve generalization compared to directly using the (hard) prediction (Hinton et al., 2015). More details to illustrate this concept can be found in the next section. $d(\cdot,\cdot)$ is a distance metric and is computed based on training samples with complete modalities. The idea for the second term is that we consider the model based on incomplete modalities, $f^{(1:m)}$, as a "student", who learns from the model based on complete modalities, $f^{(1:M)}$, as a "teacher", regarding how to make predictions for given samples. That is, we want the student model to distill knowledge from the teacher model. Thus, the second term is called a knowledge distillation (KD) loss. The third term in the optimization is to penalize model complexity. λ_1 and λ_2 are tuning parameters. To help understand the basic idea of the model in Equation (1) for modeling incomplete modalities, Figure 2 provides a graphical illustration for a simple case of two modalities.

Finally, we would like to discuss the choice of the model $f^{(1:m)}$ in the UMoS framework, $m=1,\ldots M-1$. The model needs to have the capability of quantifying the uncertainty of its prediction. To have this capability, one choice for the model $f^{(1:m)}$ is probabilistic classifiers such as logistic regression, naïve Bayes, linear and quadratic discriminant analysis, etc. A probabilistic classifier can generate a predicted probability for a patient to be in class 1, $p^{(1:m)}$, versus class 0, $1-p^{(1:m)}$, given the data modalities of the patient, $X^{(1:m)}$. This is to consider that $y|X^{(1:m)}$ follows a Bernoulli distribution with parameter $p^{(1:m)}$. It is known that the variance of a Bernoulli distribution is equivalent to Shannon entropy for uncertainty quantification (Kala, 2022). Thus, we adopt the variance of the Bernoulli distribution, $p^{(1:m)}\left(1-p^{(1:m)}\right)$, to represent the predictive uncertainty in this paper, i.e., $u^{(1:m)}=p^{(1:m)}\left(1-p^{(1:m)}\right)$. An alternative approach is to adopt a Bayesian model where the posterior distribution of $y|X^{(1:m)}$ captures the uncertainty of the prediction. A Bayesian model has the advantage of better accounting for parameter uncertainty. This is a limitation for the Bernoulli variance-based uncertainty quantification as the variance is computed based on point estimates of model parameters (i.e., estimates for minimizing a loss function as defined in Equation (1)). On the other hand, there are well-known challenges for Bayesian methods. For instance, analytical solutions for the posterior distribution may not exist. Even though approximate and computational methods can be used, these methods can be computationally complex and error-prone. Also, Bayesian methods rely on a proper selection of the prior distribution for model parameters. However, there is usually a lack of domain knowledge for the prior selection. An improper prior will not only affect the correctness of the posterior distribution but also incur computational cost. Based on these considerations, we adopt Bernoulli variance for quantifying predictive uncertainty in

this paper, while leaving the integration of a Bayesian model into the UMoS framework for future exploration.

2.3.3. Implementation and algorithm

To solve the optimization in Equation (1), we will need to choose a specific type of model for $f^{(1:m)}$. In this section, we present an implementation with $f^{(1:m)}$ as a penalized logistic regression. Logistic regression is a probabilistic classifier, thus being a proper model for UMoS according to the discussion at the end of the previous section. Let $\eta^{(1:m)} = \beta^{(1:m)} TX^{(1:m)}$ be a linear predictor that combines features contained in the collective modalities $X^{(1:m)}$ using coefficients $\beta^{(1:m)}$. Logistic regression links $\eta^{(1:m)}$ with $p^{(1:m)}$ by a logistic function, i.e., $p^{(1:m)} = \frac{1}{1+\exp\left(-\eta^{(1:m)}\right)} = \frac{1}{1+\exp\left(-\beta^{(1:m)} TX^{(1:m)}\right)}$. Then, the optimization in Eq. (1) becomes:

$$\min_{\beta^{(1:m)}} \overbrace{\sum_{i=1}^{n^{(1:m)}} l\left(y_i, p_i^{(1:m)}\right)}^{prediction\ \ loss} + \lambda_1 \overbrace{\sum_{j=1}^{n^{(1:M)}} d\left(p_j^{(1:m)}, p_j^{(1:M)}; T\right)}^{knowledge\ \ dstillain\ \ (KD)} + \lambda_2 \overbrace{\|\beta^{(1:m)}\|}^{model\ \ complexity}, \quad m = 1, \ldots M-1. \tag{2}$$

A commonly used prediction loss for logistic regression is the negative log-likelihood, which is equivalent to the cross-entropy (CE) loss (Good, 1992), i.e.,

$$l\left(y_i, p_i^{(1:m)}\right) = -\left\{y_i \log\ p_i^{(1:m)} + (1-y_i) \log\ \left(1 - p_i^{(1:m)}\right)\right\}.$$

The KD loss is defined as the KL-divergence between the "soft" predicted distributions by $\tilde{f}^{(1:m)}$ (student model) and $\tilde{f}^{(1:M)}$ (teacher model), i.e.,

$$d\left(p_j^{(1:m)}, p_j^{(1:M)}; T\right) = \tilde{p}_j^{(1:M)} \log \frac{\tilde{p}_j^{(1:M)}}{\tilde{p}_j^{(1:m)}} + \left(1 - \tilde{p}_j^{(1:M)}\right) \log \frac{\left(1 - \tilde{p}_j^{(1:M)}\right)}{\left(1 - \tilde{p}_j^{(1:m)}\right)},$$

where $\tilde{p}_j^{(1:M)}$ and $\tilde{p}_j^{(1:m)}$ denote the "soft" predicted probabilities by the teacher and student models, respectively, i.e.,

$$\tilde{p}_j^{(1:M)} = \frac{1}{1+\exp\left(-\eta^{(1:M)}/T\right)}, \tag{3}$$

$$\tilde{p}_j^{(1:m)} = \frac{1}{1+\exp\left(-\eta^{(1:m)}/T\right)}. \tag{4}$$

T is known as the temperature parameter (Hinton et al., 2015). When $T = 1$, Equation (3) and (4) become the predicted probabilities. However, using $T > 1$ has been shown to produce more generalizable results in KD. T is treated as a tuning parameter in the optimization.

The third term in Equation (2), i.e., $\|\beta^{(1:m)}\|$, can be any sparsity-inducing penalty on the model coefficients. We found that the elastic net penalty works particularly well with our dataset, due to the high-dimensional and correlated features. The elastic net penalty is:

$$\|\beta^{(1:m)}\| = \gamma\|\beta^{(1:m)}\|_1 + (1-\gamma)\|\beta^{(1:m)}\|_2^2,$$

Where $\|\cdot\|_1$ is the L1-norm and $\|\cdot\|_2^2$ is the squared L2-norm.

The optimization problem in Equation (2) is convex. To solve it, we used the adaptive moment estimation (Adam) solver (Kingma & Ba, 2014), which is a computationally-efficient version of Stochastic Gradient Descent algorithms with momentum and adaptive learning rate.

2.3.4. Hyper-parameter tuning

There are two sets of hyper-parameters to be tuned. The first set includes four tuning parameters involved in solving the optimization for the model $f^{(1:m)}$ i.e., λ_1, λ_2, T, and γ. The second set includes the uncertainty threshold associated with the model, i.e., $u^{*(1:m)}$, which is used to determine if the next modality should be acquired for each patient, $m = 1, \ldots M-1$. Since the two sets of parameters serve different functions, two validation sets, D_{val1} and D_{val2} are needed to tune each set. D_{val1} is used to select λ_1, λ_2, T, and γ for $f^{(1:m)}$. Thus, this set only needs to include samples with collective modalities $X^{(1:m)}$ available. D_{val2} is used to select the uncertainty thresholds. With different values of the uncertainty thresholds, a patient may or may not require the next modality. These different values need to be tried to find the best ones, which means that D_{val2} must include samples with complete modalities. This will further allow us to compare UMoS and the complete-modality method that requires all patients to have all/complete modalities.

The specific tuning process is the following: Based on D_{val1}, we select λ_1, λ_2, T, and γ to minimize the validation CE loss for each model $f^{(1:m)}$, $m = 1, \ldots M-1$. Then, under the selected optimal tuning parameters, the models are retrained. Next, the retrained models are used in the UMoS workflow with a grid search for the uncertainty thresholds. That is, for each combination of values of the uncertainty thresholds, the workflow is deployed on samples in D_{val2}. After the deployment, we can compute two types of metrics on D_{val2} : 1) Area Under the Curve (AUC), which reflects the classification performance of UMoS. AUC is chosen because it is not affected by the probability cutoff used to classify samples, whereas other classification metrics can be reported too. 2) Percentages of samples that need to acquire $X^{(1)}$, $X^{(1:2)}$, …, $X^{(1:M-1)}$, respectively, which reflect the data efficiency of UMoS, i.e., the percentages of patients that can be saved from needing to acquire more expensive/less accessible modalities. We will cross-reference the metrics in 1) and 2) to choose the uncertainty thresholds that yield a high AUC while at the same time saving as many as possible

patients from needing more expensive/less accessible modalities. In practice, instead of providing a single choice for the uncertainty thresholds, it may be more desirable to create a visualization to show the tradeoff between the metrics in 1) and 2) for a range of different choices. This will help practitioners choose the uncertainty thresholds that best suit their needs. More details to demonstrate the procedure of selecting the uncertainty thresholds will be presented in the case studies in Sec. 3.

Furthermore, it is worth mentioning that an alternative approach to using two validation sets D_{val1} and D_{val2} is to use a double-loop cross-validation (CV) scheme, which includes an internal CV serving the role of D_{val1} and an external CV serving the role of D_{val2}. CV provides a more robust approach for hyper-parameter tuning as it iterates through all samples, which is used in our case study.

Finally, we want to point out that the selection of uncertainty thresholds as described above does not consider the available resources such as the availability of imaging equipment and appointment slots. In other words, UMoS makes suggestions as to whether a patient needs a subsequent modality such as an amyloid-PET scan. Whether and when the patient can acquire that modality will have to depend on the available resources in the health care system.

2.4. Theoretical study

This section discusses some theoretical aspects of the models included in the UMoS frameworks, $f^{(1:m)}$, $m = 1, \ldots M-1$. For notation simplicity, we remove the superscript and use f to represent any model in this sequence. The other notations used in this section are defined as follows: Let x denote the input feature set. Let $y = 1$ *or* 0 denote two classes. Recall that f is a probabilistic classifier. Thus, given x, $f(x)$ outputs the predicted probability for x to be class 1. To get the predicted class label $\hat{y}$, a cutoff of 0.5 is typically used and $\hat{y} = \mathbb{I}(f(x) > 0.5)$, where $\mathbb{I}(\cdot)$ is an indicator function. Furthermore, considering that the predicted class is binary and can be modeled by a Bernoulli distribution, the variance of the distribution is $f(x)(1-f(x))$, which can represent the predictive uncertainty, i.e., $u(x) = f(x)(1-f(x))$. Also, let $a(x)$ denote the probability of correct prediction/classification for x, i.e., $a(x) = P(\hat{y} = y|x)$. Finally, let $\phi(f)$ denote a loss for f. Let $R(f) = \mathbb{E}_{x,\ y}(\ \phi(f))$ denote the risk associated with the loss ϕ, which is the expected loss on the data distribution. Let $\hat{R}(f)$ denote the empirical risk computed based on a training set. Let $R^* = inf_f R(f)$ denote the Bayesian optimal risk. For a given f, the excess risk is $R(f) - R^*$.

Different from typical classifiers that aim to maximize accuracy, f needs to consider both accuracy and uncertainty. Specifically, f needs to be trained such that if it is more certain about a prediction, the probability that the prediction is correct should be higher (i.e., the prediction is more accurate). This is named the more-certain-more-accurate (MCMA) condition in this paper. Definition 1 provides a formal definition of MCMA.

Definition 1 (MCMA condition): Consider two samples x and x' which are predicted by the model f. If $u(x) < u(x')$, i.e., the model is more certain about the prediction for x, then $a(x) > a(x')$, i.e., the probability for x to be correctly predicted/classified is also higher.

The training of f should be attentive to the MCMA condition because f is not a stand-alone classifier. In UMoS, the predictive uncertainty (or certainty) of f for a given patient will trigger the decision as to whether the next data modality needs to be acquired. If the decision is not to acquire the next modality for a patient, which happens when the certainty of the prediction is high, we want the probability that the prediction is correct for this patient to also be high. In other words, we want to avoid training a model that has high certainty for its prediction but the prediction result is actually wrong, because this would stop the patient from acquiring another modality to improve the prediction accuracy.

To train a model f with the aforementioned property, we first need to define a loss that encodes the MCMA condition, which is the MCMA loss as follows.

Definition 2 (MCMA loss): The MCMA loss is defined on a pair of samples with $u(x) < u(x')$, i.e., $\phi_{MCMA}(f) = \mathbb{I}(u(x) < u(x'))\mathbb{I}(a(x) \leq a(x'))$. A loss of one is incurred if the MCMA condition is violated and zero if it is satisfied.

Furthermore, we can write the MCMA risk as $R_{MCMA}(f) = \mathbb{E}_{x,y}(\phi_{MCMA}(f))$. To train a classifier, the risk cannot be directly minimized as the data distribution is unknown. Thus, a typical training process is to minimize the empirical risk computed based on a training set. Specifically, given a training set of n sample, $\mathcal{D} = (x_i, y_i)_{i=1}^n$, the empirical MCMA risk is $\hat{R}_{MCMA}(f) = \frac{1}{n(n-1)/2}\sum_{(i,j)\in\mathcal{D};\ u(x_i)<u(x_j)}\mathbb{I}\big(u(x_i) < u(x_j)\big)\mathbb{I}(a(x_i) \leq a(x_j))$, where the summation is over all pairs of samples in the training set and there are a total of $n(n-1)/2$ pairs.

It is difficult to minimize the empirical MCMA risk in model training because the risk is intractable in optimization. In Theorem 1, we derive that minimizing MCMA risk is equivalent to minimizing the 0/1 risk which is well-defined for classification problems.

Theorem 1: *The 0/1 loss is* $\phi_{0/1}(f) = \mathbb{I}\big(\hat{y} \neq y|x\big)$, *i.e., a loss of one is incurred for wrong prediction/classification. The 0/1 risk is* $R_{0/1}(f) = \mathbb{E}_{x,\ y}\big(\phi_{0/1}(f)\big)$. *The minimizer for the MCMA risk is the same as that for the 0/1 risk, i.e., arg* $inf_f\ R_{MCMA}(f) = \arg\inf_f\ R_{0/1}(f)$.

Proof: We first prove $u(x) = a(x)(1-a(x))$. Recall that, according to the definition of $u(x)$, $u(x) = f(x)(1-f(x))$, where $f(x)$ outputs the predicted probability for x to be class 1. $f(x)$ can be written in a probabilistic form, $f(x) = P(\hat{y} = 1|x)$. Put this back into $u(x)$, we get

$$u(x) = P\big(\hat{y} = 1|x\big)\big(1 - P\big(\hat{y} = 1|x\big)\big). \tag{5}$$

When the true class is $y = 1$, we can use y to replace the 1 in (5). Thus, (5) becomes $u(x) = P(\hat{y} = yx)(1 - P(\hat{y} = yx)) = a(x)(1-a(x))$. When the true class is $y = 0$, we first write (5) into an equivalent form, i.e., $u(x) = \big(1 - P(\hat{y} = 0|x)\big)\big(P(\hat{y} =$

$0|x)$), and then replace the 0 by y, so that we get $u(x) = \big(1 - P(\hat{y} = y|x)\big)\big(P(\hat{y} = y|x)\big) = (1 - a(x))a(x)$. Thus, we proved that no matter if $y = 1$ or 0, $u(x) = a(x)(1 - a(x))$. Furthermore, using the found relationship between u and a, we can show that, for two samples with $u(x) < u(x')$, a sufficient and necessary condition for $a(x) \leq a(x')$ is $a(x) \leq 0.5$. This means that the MCMA loss is incurred if and only if $a(x) \leq 0.5$. Recall that $a(x)$ is the probability of correct prediction, so $a(x) \leq 0.5$ means the prediction is wrong. When this happens, the 0/1 loss is incurred. Thus, we have proved that the MCMA loss is incurred if and only if the 0/1 loss is incurred. Furthermore, since risks are expected losses on the data distribution, we can naturally show that minimizing the MCMA risk is equivalent to minimizing the 0/1 risk, i.e., $arg\ inf_f\ R_{MCMA}(f) = \arg\inf_f R_{0/1}(f)$. ■

The 0/1 loss is a discrete loss, which makes the corresponding 0/1 risk difficult to optimize in model training. We look for a surrogate loss of the 0/1 loss, which has better computational tractability. Lemma 1 provides an upper bound of the excess risk associated with the 0/1 loss (See Theorem 1 in Bartlett et al. (2006)). Based on the result of Lemma 1, we further derive in Theorem 2 that the CE loss is a convex, surrogate of the 0/1 loss.

Lemma 1: *For a given classifier f, the excess 0/1 risk is $R_{0/1}(f) - R^*_{0/1}$. Let $\phi_s(f)$ denote a general s-loss that is convex and classification-calibrated (see Definition 1 in Bartlett et al. (2006)), and the associated excess risk is $R_s(f) - R^*_s$. Then, for any classifier f, we have $\psi\left(R_{0/1}(f) - R^*_{0/1}\right) \leq R_s(f) - R^*_s$, where $\psi(\theta) = H^-_s\left(\frac{1+\theta}{2}\right) - H_s\left(\frac{1+\theta}{2}\right)$ is a transform function, $\psi : [0,\ 1] \to [0,\ \infty)$.*

Theorem 2: *For any sequence of classifiers $f^{(n)}$ under the CE loss ϕ_{CE}, if $R_{CE}\left(f^{(n)}\right) \to R^*_{CE}$, then $R_{0/1}\left(f^{(n)}\right) \to R^*_{0/1}$, which means that convergence of the CE risk also leads to convergence of the 0/1 risk.*

Proof: Under the CE loss, the transform function in Lemma 1 can be derived as $\psi(\theta) = H^-_{CE}\left(\frac{1+\theta}{2}\right) - H_{CE}\left(\frac{1+\theta}{2}\right) = \frac{1+\theta}{2}\ln(1+\theta) + \frac{1-\theta}{2}\ln(1-\theta)$, which is a non-negative monotonically increasing function on $\theta \in [0, 1]$. Convergence of the CE risk means that there exists an integer N such that $R_{CE}\left(f^{(N)}\right) - R^*_{CE}(f) \leq \psi(\varepsilon)$ for all ε. Furthermore, it can be shown that the CE loss is convex and classification-calibrated. Thus, we can use the result of Lemma 1 and get $\psi\left(R_{0/1}\left(f^{(N)}\right) - R^*_{0/1}\right) \leq R_{CE}\left(f^{(N)}\right) - R^*_{CE}(f) \leq \psi(\varepsilon)$. Recall that the function ψ is non-negative and monotonically increasing. Thus, $R_{0/1}\left(f^{(N)}\right) - R^*_{0/1} \leq \varepsilon$, which means that the 0/1 risk converges. ■

Remarks: In Lemma 1, according to Definition 1 in Bartlett et al. (2006), "convexity" means that the loss function is convex with respect to the predicted probability of the classifier, not with respect to the parameters/coefficients of the classifier. "Classification-calibration" means that the loss function is defined in a meaningful way that it penalizes wrong predictions while not penalizing correct predictions. The CE loss satisfies these two properties, which leads to the result in Theorem 2.

Theorem 2 implies that the CE loss is a surrogate of the 0/1 loss, and minimization of the CE risk can lead to minimization of the 0/1 risk. Also, based on Theorem 1, we know that the minimizer for the MCMA risk is the same as that for the 0/1 risk. Thus, the minimization of the CE risk can lead to minimization of the MCMA risk. It is important that the models included in the UMoS framework should respect the MCMA condition and that the training of these models aims to minimize the MCMA risk. Through the theoretical study in this section, we demonstrated that this important goal can be achieved by using the CE loss.

Another implication of the theoretical results in this section is that we can potentially include any classifier that is trained using the CE loss in the UMoS framework. In the case studies of this paper, we demonstrate the UMoS framework using a penalized logistic regression model that is estimated with the CE loss. Logistic regression is adopted due to its simplicity, interpretability, and ease in training. Other classifiers can be adopted as the CE loss is a commonly used loss function in training classifiers. Also, future work may include designing other surrogate losses beyond CE that can provide a tighter bound of the excess 0/1 risk, thus leading to better performance for supporting the MCMA condition.

3. Result

3.1. UMoS with two modalities

We first demonstrated the application of the UMoS framework to the ADNI dataset under a two-modality scenario: demographic/clinical data and MRI together as the first modality $X^{(1)}$; amyloid-PET as the second modality $X^{(2)}$. MRI is currently used in the standard of care in the U.S. for AD-related clinical examination. Thus, it is reasonable to assume that MRI and demographic/clinical data have the same accessibility to patients. In comparison, amyloid-PET has much lower accessibility due to the high cost.

To apply UMoS, two models were trained: $f^{(1)}$ based on $X^{(1)}$, which is considered an incomplete modality; $f^{(1:2)}$ based on complete modalities, $X^{(1:2)}$. As mentioned in Section 2.3.2, the model based on complete modalities, $f^{(1:2)}$, can be trained using any classification algorithm. We tried a variety of algorithms and found logistic regression with the elastic net penalty yielded the best performance in our experiments. Furthermore, we trained the model based on the incomplete modality, $f^{(1)}$, using the proposed model in Equation (2). A double-loop 10-fold CV scheme was used for hyper-parameter tuning of UMoS.

Figure 3 shows the AUC of UMoS evaluated on external CV under different values of the uncertainty threshold $u^{*(1)}$.

As $u^{*(1)}$ increases, more patients will be saved from needing to acquire amyloid-PET, but with the price of lowering the AUC. To select a proper value for $u^{*(1)}$, we started with $u^{*(1)} = 0.028$, which yielded the highest AUC = 0.905. This setting has zero data efficiency because it requires all patients to acquire amyloid-PET in addition to demographic/clinical data and MRI. Then, we gradually increased $u^{*(1)}$ until the AUC decreased to a level below which the AUC would become significantly lower than 0.905. We adopted a conservative approach and set the significance level to be *p*-value equal to 0.2. Under this significance level, we could decrease the AUC to 0.88, which has no statistically significant difference from 0.905. The corresponding uncertainty threshold is $u^{*(1)} = 0.237$. This setting comes with a huge gain in data efficiency that 77% patients can be saved from needing to acquire amyloid-PET (i.e., they only need to acquire demographic/clinical data and MRI). Using this setting, Figure 4 demonstrates the UMoS workflow and reports metrics about the classification performance and data efficiency associated with the workflow. Furthermore, Table 1 compares UMoS and the complete-modality method (i.e., the method that requires all patients to acquire complete modalities) in terms of AUC, accuracy (ACC), sensitivity, specificity, and data efficiency. The *p*-value of the comparison were calculated using the following approaches: The *p*-value of ACC, sensitivity and specificity were calculated using McNemar's test based on the confusion matrix constructed from comparing the true class of each sample and the corresponding predicted class from CV (Hofer et al., 2020). The *p*-value of AUC was calculated using Delong's test (Sun & Xu, 2014). The *p*-value of data efficiency was calculated using the z-test for proportions. Compared to the complete-modality method, UMoS has no statistically significant difference for AUC, ACC, sensitivity, and specificity, but with a significant gain in data efficiency.

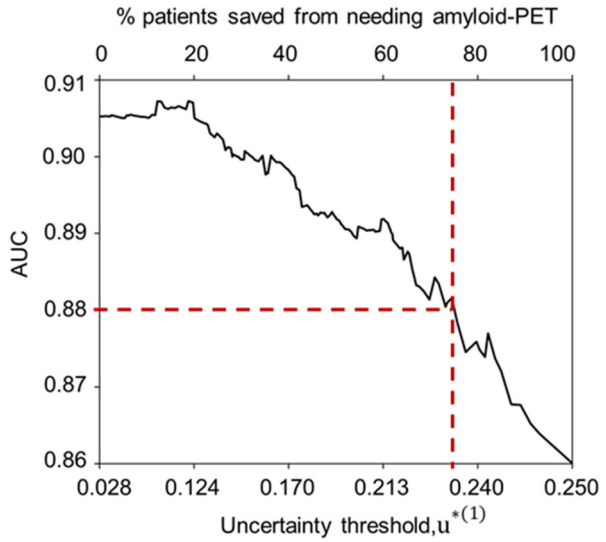

Figure 3. The change of AUC (y-axis) with respect to uncertainty threshold $u^{*(1)}$ (x-axis, bottom). Each uncertainty threshold (x-axis, bottom) corresponds to a percentage of patients who are saved from needing to acquire amyloid-PET (x-axis, top). Intersection of two red dash lines marks $u^{*(1)} = 0.237$, which results in AUC = 0.88 (no significant difference from the highest AUC) and 77% patients saved from needing amyloid-PET.

Furthermore, to demonstrate the effectiveness of the uncertainty threshold selection strategy in UMoS, we randomly selected 77% patients to have no amyloid-PET, referred to as the random strategy. We compared the AUCs of the two strategies based on non-overlapping patients selected by these strategies. The average AUC over ten repeated runs of the random strategy is 0.796. This is lower than the AUC of the proposed strategy in UMoS, which is 0.847.

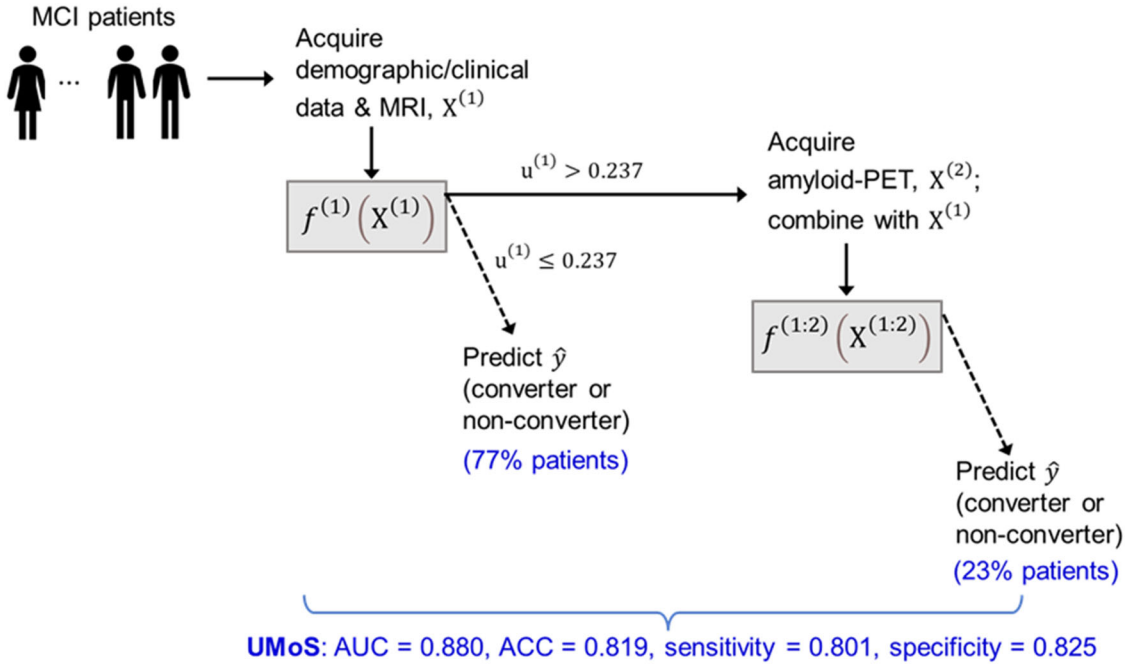

Figure 4. UMoS workflow and performance metrics for the two-modality scenario. Using an uncertainty threshold of 0.237, 77% patients only need to acquire demographic/clinical data and MRI while 23% patients need to acquire all the modalities.

Table 1. Comparison between UMoS and the complete-modality method (i.e., the method that requires all patients to acquire complete modalities).

	AUC	ACC	Sensitivity	Specificity	Data efficiency (% patients saved from needing amyloid-PET)
UMoS	0.880	0.819	0.801	0.825	77%
Compete-modality method	0.905	0.830	0.840	0.827	0%
P value of difference	0.2	0.5	0.2	1.0	<0.001

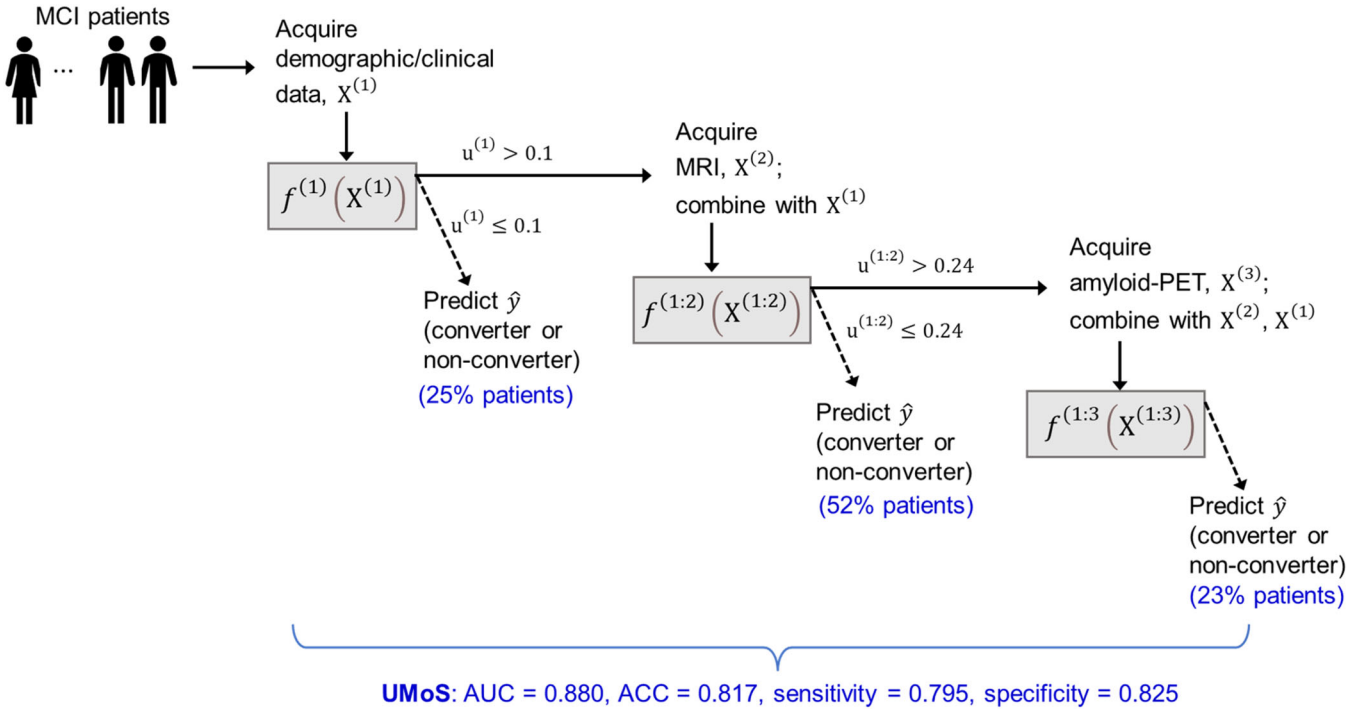

Figure 5. UMoS workflow and performance metrics for the three-modality scenario. Using two uncertainty thresholds of 0.1 and 0.24, 25% patients only need demographic/clinical data, 52% patients need both demographic/clinical data and MRI, and 23% patients need all three modalities.

Moreover, to demonstrate the unnecessity of adding amyloid-PET for the 77% patients selected by UMoS, we applied the complete-modality model to these patients, which yielded an AUC of 0.923. This has no significant difference from the AUC of 0.901 by the incomplete-modality model ($p = 0.3$). Also, to demonstrate the necessity of adding amyloid-PET for the remaining 23% patients, we applied the incomplete-modality model to these patients, which yielded an AUC of 0.587. This is much lower than that from the complete-modality model which achieved an AUC of 0.793.

Finally, we would like to point out that all the aforementioned results are based on an uncertainty threshold, $u^{*(1)} = 0.237$, which is chosen such that the AUC of UMoS is not significantly lower than that based on the complete-modality model at a significance level of 0.2. This significance level is used as a conservative choice. More stringent significance levels can be used, which will yield a lower AUC but a higher percentage of patients saved from needing amyloid-PET. For example, at a significance level of 0.05, the AUC of UMoS is 0.866 with 91% patients saved from needing amyloid-PET.

3.2. UMoS with three modalities

In this section, we demonstrate the application of UMoS to the ADNI dataset under a three-modality scenario: demographic/clinical data as the first modality $X^{(1)}$; MRI as the second modality $X^{(2)}$; amyloid-PET as the third modality $X^{(3)}$. Even though MRI is considered an appropriate standard of care by medical specialists in AD-related examinations in the U.S., primary care providers may take a more conservative approach. Also, if patients do not have insurance coverage, the cost of MRI is much higher than that of obtaining basic demographic/clinical data. Thus, in this section, as a proof-of-concept, we apply UMoS to sequentially add the three modalities on an as-needed basis for each patient.

The training process is similar to that in Section 3.1. We report the results in the following. There are two uncertainty thresholds, $u^{*(1)}$ and $u^{*(1:2)}$. The highest AUC = 0.905 is achieved with $u^{*(1)} = 0.005$ and $u^{*(1:2)} = 0.042$. This setting, however, has zero data efficiency because it requires all patients to acquire all data modalities. If the two uncertainty thresholds are increased to $u^{*(1)} = 0.1$ and $u^{*(1:2)} = 0.24$, AUC = 0.88 will be achieved without a significant difference from the highest AUC ($p = 0.2$). This setting improves the data efficiency so that 77% patients can be exempted from the need to acquire amyloid-PET (i.e., they only need to acquire demographic/clinical data and MRI); 25% patients can be exempted from the need to acquire MRI and amyloid-PET (i.e., they only need to acquire demographic/clinical data). Figure 5 demonstrates the UMoS workflow and reports metrics about the classification performance and

Table 2. Comparison between UMoS and the complete-modality method (i.e., the method that requires all patients to acquire complete modalities).

	AUC	ACC	Sensitivity	Specificity	Data efficiency (% patients saved from needing amyloid-PET)	Data efficiency (% patients saved from needing MRI and amyloid-PET)
UMoS	0.880	0.817	0.795	0.825	77%	25%
Complete-modality method	0.905	0.830	0.840	0.827	0%	0%
P-value of difference	0.2	0.4	0.1	1.0	<0.001	<0.001

data efficiency associated with this workflow. Furthermore, Table 2 compares UMoS and the complete-modality method (i.e., the method that requires all patients to acquire complete modalities) in terms of AUC, ACC, sensitivity, specificity, and data efficiency.

To demonstrate the unnecessity of adding subsequent modalities for those patients selected by UMoS, we performed the following experiments. Among the patients who only need demographic/clinical data according to UMoS, we compared the AUCs of the model using only demographic/-clinical data and the model that added MRI. The AUCs of the two models are 0.962 and 0.967, which do not have significant difference ($p = 0.8$). Furthermore, among the patients who need demographic/clinical data and MRI according to UMoS, we compared the AUCs of the model using demographic/clinical data and MRI with the model that added amyloid-PET. The AUCs of the two models are 0.835 and 0.884, which do not have significant difference ($p = 0.2$).

Also, to demonstrate the necessity of adding subsequent modalities, we performed the following experiments. Among the patients who need both demographic/clinical data and MRI according to UMoS, we compared the AUCs of the model using only demographic/clinical data and the model that added MRI. The AUC of the former is 0.794, while adding MRI increases the AUC to 0.835. Furthermore, among the patients who need all three modalities according to UMoS, we compared the AUCs of the model using only demographic/clinical data and MRI with the model that added amyloid-PET. The AUC of the former is 0.603, while adding amyloid-PET increases the AUC to 0.781.

Finally, we would like to point out that all the aforementioned results are based on uncertainty thresholds, $u^{*(1)} = 0.1$ and $u^{*(1:2)} = 0.24$, which are chosen such that the AUC of UMoS is not significantly lower than that based on the complete-modality model at a significance level of 0.2. This significance level is used as a conservative choice. More stringent significance levels can be used, which will yield a lower AUC but a higher percentage of patients saved from needing subsequent modalities. For example, at a significance level of 0.05, the AUC of UMoS is 0.862 with 28% patients saved from needing MRI and amyloid-PET and 83% patients saved from needing amyloid PET.

4. Discussion and conclusion

We proposed an ML framework, UMoS, to predict MCI conversion to AD by sequentially adding data modalities for each patient on an as-needed basis. The capability of using fewer data modalities while preserving prediction accuracy distinguished UMoS from existing multi-modality research. This capability is also important for improving clinical efficiency especially when there are barriers preventing sophisticated modalities, which are usually more costly and less accessible, from routine use. We applied UMoS to an ADNI dataset. In the two-modality case, we demonstrated that 77% of patients can be saved from needing to acquire PET, whereas the prediction accuracy has no significant difference from the ML model based on all modalities (0.880 AUC compared to 0.905 AUC, p-value of difference = 0.2). We further demonstrate in a three-modality scenario that if MRI and demographic/clinical data were split into two modalities, 25% of patients can be saved from needing MRI without impacting the prediction accuracy. These results show the high accuracy and data efficiency achieved by UMoS.

To our best knowledge, there is no existing study for personalized modality saving in predicting MCI conversion to AD. In our experiments, we showed that the proposed UMoS framework can preserve prediction accuracy at the same level as that achieved by ML models based on all/complete modalities. In the existing studies of predicting MCI conversation to AD using multi-modality datasets, the reported accuracy is in the range of 0.743–0.898 by integrating MRI and PET with some works additionally including non-imaging data (Liu et al., 2017; Liu et al., 2021; Shen et al., 2021; Xu et al., 2016; Zhang & Shi, 2020; Zhou et al., 2019; Zhou et al., 2020; Zhou et al., 2019; Zhu et al., 2019). UMoS achieved similar levels of accuracy, but with a significant saving of data modalities. On the other hand, we want to point out that the different studies vary in aspects of dataset composition, MCI conversion timeframe, and features included, and thus a direct comparison is difficult. To mitigate the difference in results caused by factors other than the ML model, we used standard, well-established image processing and feature extraction methods from MRI and PET, adopted a commonly used MCI conversion timeframe (36 months), and utilized data from the world-class ADNI database. Comparative prediction performance was achieved using the proposed UMoS framework with significantly improved data efficiency, which indicated the utility of UMoS in this research field.

AD is a devastating neurodegenerative disease. Early prediction of AD can lead to early intervention to potentially slow down the disease's progression. However, in previous studies, the cost and accessibility of the different data modalities used to obtain such early prediction did not raise enough concern in developing ML-driven computer-aided diagnosis/prognosis systems. The tracking of the progression of AD can be a long-term and costly process for patients. The proposed UMoS framework provided a clinical tool to

assist with deciding what data modalities/diagnostic exams each patient needs, and a modality/exam is added only if the patient needs it. In this way, some patients can be saved from needing all the modalities while the prediction accuracy for these patients would not be compromised, thus lessening the burden on patients and the healthcare system.

To gain some insight into patients with which characteristics are more likely to be exempted from MRI or PET as informed by UMoS, we performed some additional experiments. Specifically, in one experiment we treated the patients who are exempted from MRI and who are not as two classes, and built a decision tree model to classify them based on demographic and clinical characteristics of the patients. We focused on patients who are converters because the non-converters are heterogenous and include patients who converted to AD in any number of years beyond three years as well as patients whose MCI symptoms are not due to AD as the underlying etiology. Our result showed that patients who can be exempted from MRI have older age and higher CDR (meaning worse symptoms). A similar result was obtained in another experiment which aimed to find out patients with which characteristics are more likely to be exempted from PET. These results make sense because patients with these characteristics, by nature, have a higher chance of developing AD. Thus, it is more likely for UMoS to find the predictions for these patients to be certain even without imaging exams.

This study has some limitations, which drive future research directions. First, an assumption of UMoS is that every patient should have access to the first modality, such as the demographic/clinical data and MRI in the two-modality case study. This assumption is met in our study because we chose to include commonly used demographic and clinical variables; MRI is also commonly available because it is used in the standard of care in the U.S. for AD-related clinical examinations. However, there could be situations when this assumption cannot be satisfied. For example, the demographic/clinical data or/and MRI for some patients may be missing due to problems that happen during the process of data acquisition, storage, and preprocessing; the MRI scans of some patients may not have sufficient quality to be used by ML. Thus, there could be patients who do not have demographic/clinical data or/and MRI but only amyloid-PET. To include the data of such patients in training the UMoS model, one potential approach is to impute the missing modalities. Future research may be conducted to design proper imputation algorithms to be integrated with UMoS and to assess the uncertainty of modality selection due to the incorporation of imputed data in training the UMoS.

Second, the proposed sequential modality selection process can be extended to include more and other data modalities, such as genomic data, CSF biomarkers, and other neuroimaging techniques (fMRI, tau-PET, etc.). Early prediction of AD is a challenging task. These data modalities have shown utility in predicting MCI conversion to AD. An extended UMoS would help with selecting the needed modalities for each patient to save costs.

Third, there are several aspects of the models in UMoS that can be improved. The uncertainty threshold of each model is currently determined by cross-validation. An integrated approach may work better, which learns the threshold together with model coefficients during the training process. The models are based on pre-extracted features from anatomically-defined brain regions. Future research may extend this work and use deep learning models based on 3D images. There is heterogeneity in the AD risk among patients, depending on patient characteristics. Even though we included some patient characteristic variables which are known AD risk factors in our present model, a more carefully-crafted method is to stratify patients into cohorts with different characteristic profiles and train a different model to link imaging data with AD conversion status, and identify a different uncertainty threshold for each cohort. This method can be explored in future research. Last but not least, this study focused on prediction of MCI conversion to AD. Prediction of future risk of AD for individuals who are currently cognitive normal is important to move the detection window even earlier, i.e., before MCI, which can lead to significant improvement in the treatment potential of this devastating disease.

Funding

This research was primarily supported by NIH grant 2R42AG053149-02A1 and NSF grant DMS-2053170. This research was also supported by NIH grants RF1AG073424, R01AG069453 and P30AG072980, the State of Arizona, and Banner Alzheimer's 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. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Consent and approval statement

This study has been exempted from the requirement for approval by an institutional review board. The data corpus is publicly available.

Disclosure statement

The authors report no conflict of interest.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Alzheimer's, A. (2022). 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *18*(4), 700–789. https://doi.org/10.1002/alz.12638

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156. https://doi.org/10.1198/016214505000000907

Beheshti, I., Demirel, H., & Matsuda, H. (2017). Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Computers in Biology and Medicine*, 83, 109–119.

Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., Aalten, P., De Deyn, P. P., Biessels, G. J., Claassen, J. A. H. R., Middelkoop, H. A. M., Smits, M., Niessen, W. J., van Swieten, J. C., van der Flier, W. M., Ramakers, I. H. G. B., & van der Lugt, A. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage. Clinical*, *31*, 102712. https://doi.org/10.1016/j.nicl.2021.102712

Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M. E., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J. R., Garrett, C., Bastos-Leite, A. J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., … Klein, S. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage*, *111*, 562–579. https://doi.org/10.1016/j.neuroimage.2015.01.048

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/s0896-6273(02)00569-x

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., & Dale, A. M. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *14*(1), 11–22. https://doi.org/10.1093/cercor/bhg087

Good, I. J. (1992). *Rational decisions (Breakthroughs in statistics* (pp. 365–377). Springer.

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), 1–16. https://doi.org/10.1186/s40537-019-0206-3

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).

Hofer, I. S., Lee, C., Gabel, E., Baldi, P., & Cannesson, M. (2020). Development and validation of a deep neural network model to predict postoperative mortality, acute kidney injury, and reintubation using a single feature set. *NPJ Digital Medicine*, *3*(1), 58. https://doi.org/10.1038/s41746-020-0248-0

Holtzman, D. M., Morris, J. C., & Goate, A. M. (2011). Alzheimer's disease: The challenge of the second century. *Science Translational Medicine*, *3*(77), 77sr71–77sr71. https://doi.org/10.1126/scitranslmed.3002369

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Kala, Z. (2022). Quantification of Model Uncertainty Based on Variance and Entropy of Bernoulli Distribution. *Mathematics*, *10*(21), 3980. https://doi.org/10.3390/math10213980

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Liu, M., Zhang, J., Yap, P.-T., & Shen, D. (2017). View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. *Medical Image Analysis*, *36*, 123–134. https://doi.org/10.1016/j.media.2016.11.002

Liu, X., Chen, K., Weidman, D., Wu, T., Lure, F., & Li, J. (2021). A novel transfer learning model for predictive analytics using incomplete multimodality data. *IISE Transactions*, *53*(9), 1010–1022. https://doi.org/10.1080/24725854.2020.1798569

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002

Rosenberg, P. B., Wong, D. F., Edell, S. L., Ross, J. S., Joshi, A. D., Brašić, J. R., Zhou, Y., Raymont, V., Kumar, A., Ravert, H. T., Dannals, R. F., Pontecorvo, M. J., Skovronsky, D. M., & Lyketsos, C. G. (2013). Cognition and amyloid load in Alzheimer disease imaged with florbetapir F 18 (AV-45) positron emission tomography. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, *21*(3), 272–278. https://doi.org/10.1016/j.jagp.2012.11.016

Shen, H. T., Zhu, X., Zhang, Z., Wang, S.-H., Chen, Y., Xu, X., & Shao, J. (2021). Heterogeneous data fusion for predicting mild cognitive impairment conversion. *Information Fusion*, *66*, 54–63. https://doi.org/10.1016/j.inffus.2020.08.023

Su, Y., Blazey, T. M., Snyder, A. Z., Raichle, M. E., Marcus, D. S., Ances, B. M., Bateman, R. J., Cairns, N. J., Aldea, P., Cash, L., Christensen, J. J., Friedrichsen, K., Hornbeck, R. C., Farrar, A. M., Owen, C. J., Mayeux, R., Brickman, A. M., Klunk, W., Price, J. C., … Benzinger, T. L. S. (2015). Partial volume correction in quantitative amyloid imaging. *NeuroImage*, *107*, 55–64. https://doi.org/10.1016/j.neuroimage.2014.11.058

Su, Y., D'Angelo, G. M., Vlassenko, A. G., Zhou, G., Snyder, A. Z., Marcus, D. S., Blazey, T. M., Christensen, J. J., Vora, S., Morris, J. C., Mintun, M. A., & Benzinger, T. L. S. (2013). Quantitative analysis of PiB-PET with freesurfer ROIs. *PloS One*, *8*(11), e73377. https://doi.org/10.1371/journal.pone.0073377

Sun, X., & Xu, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, *21*(11), 1389–1393. https://doi.org/10.1109/LSP.2014.2337313

Wang, Q., Zhan, L., Thompson, P., & Zhou, J. (2020). *Multimodal learning with incomplete modalities by knowledge distillation* [Paper presentation]. *Proceedings of* [Paper presentation].The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, In (pp. 1828–1838). https://doi.org/10.1145/3394486.3403234

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J., & Alzheimer's, D. N. I. (2014). Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, *102*, 192–206. https://doi.org/10.1016/j.neuroimage.2013.08.015

Xu, L., Wu, X., Li, R., Chen, K., Long, Z., Zhang, J., Guo, X., & Yao, L. (2016). Prediction of progressive mild cognitive impairment by multi-modal neuroimaging biomarkers. *Journal of Alzheimer's Disease: JAD*, *51*(4), 1045–1056. https://doi.org/10.3233/JAD-151010

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., & Ye, J. (2012). *Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. Proceedings of* [Paper presentation].The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 1149–1157). https://doi.org/10.1145/2339530.2339710

Zhang, J., Zheng, B., Gao, A., Feng, X., Liang, D., & Long, X. (2021). A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. *Magnetic Resonance Imaging*, *78*, 119–126. https://doi.org/10.1016/j.mri.2021.02.001

Zhang, T., & Shi, M. (2020). Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease. *Journal of Neuroscience Methods*, *341*, 108795. https://doi.org/10.1016/j.jneumeth.2020.108795

Zhang, Y., & Lee, A. A. (2019). Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science*, *10*(35), 8154–8163. https://doi.org/10.1039/c9sc00616h

Zhou, T., Liu, M., Thung, K.-H., & Shen, D. (2019). Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. *IEEE Transactions on Medical Imaging*, *38*(10), 2411–2422. https://doi.org/10.1109/TMI.2019.2913158

Zhou, T., Thung, K.-H., Liu, M., Shi, F., Zhang, C., & Shen, D. (2020). Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data. *Medical Image Analysis*, *60*, 101630. https://doi.org/10.1016/j.media.2019.101630

Zhou, T., Thung, K. H., Zhu, X., & Shen, D. (2019). Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human Brain Mapping*, *40*(3), 1001–1016. https://doi.org/10.1002/hbm.24428

Zhu, Q., Yuan, N., Huang, J., Hao, X., & Zhang, D. (2019). Multi-modal AD classification via self-paced latent correlation analysis. *Neurocomputing*, *355*, 143–154. https://doi.org/10.1016/j.neucom.2019.04.066