Received: 5 October 2021 | Revised: 11 April 2022 | Accepted: 10 May 2022

DOI: 10.1002/sim.9442

RESEARCH ARTICLE

Statistics in Medicine WILEY

A classification for complex imbalanced data in disease screening and early diagnosis

Yiming Li[1] | **Wei-Wen Hsu**[2] | **for the Alzheimer's Disease Neuroimaging Initiative**

[1]Department of Statistics, Kansas State University, Manhattan, Kansas, USA

[2]Division of Biostatistics and Bioinformatics, Department of Environmental and Public Health Sciences, University of Cincinnati, Cincinnati, Ohio, USA

Correspondence
Wei-Wen Hsu, Division of Biostatistics and Bioinformatics, Department of Environmental and Public Health Sciences, University of Cincinnati, Cincinnati, OH 45267, USA.
Email: hsuwe@uc.edu

Imbalanced classification has drawn considerable attention in the statistics and machine learning literature. Typically, traditional classification methods often perform poorly when a severely skewed class distribution is observed, not to mention under a high-dimensional longitudinal data structure. Given the ubiquity of big data in modern health research, it is expected that imbalanced classification in disease diagnosis may encounter an additional level of difficulty that is imposed by such a complex data structure. In this article, we propose a nonparametric classification approach for imbalanced data in longitudinal and high-dimensional settings. Technically, the functional principal component analysis is first applied for feature extraction under the longitudinal structure. The univariate exponential loss function coupled with group LASSO penalty is then adopted into the classification procedure in high-dimensional settings. Along with a good improvement in imbalanced classification, our approach provides a meaningful feature selection for interpretation while enjoying a remarkably lower computational complexity. The proposed method is illustrated on the real data application of Alzheimer's disease early detection and its empirical performance in finite sample size is extensively evaluated by simulations.

KEYWORDS

Alzheimer's disease, AUC, brain imaging data, class imbalance, group LASSO

1 | INTRODUCTION

In many disease screening and early diagnosis studies, imbalanced classification is the most common challenge when a severely skewed class distribution in the data is attributed to the rarity of the disease. Traditional classification methods, such as logistic regression and machine learning models, that generally assume a balanced class distribution often perform poorly and misclassify subjects from the minority class (ie, disease) as ones from the majority (ie, health), resulting in a high false negative rate. Although it is possible to achieve a high predictive accuracy as well as a good specificity, the sensitivity is anticipated to be low due to the high false negative rate. Imbalanced classification is even more challenging when the real data structure is complex, for example, in high-dimensional longitudinal settings. In many biomedical studies, high-dimensional longitudinal data are often collected irregularly and sparsely, where the high-dimensional measurements on each subject are taken repeatedly at discrete random time points and the number of measurements

may vary between subjects. As a good example, in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, magnetic resonance imaging (MRI) data which are generally high-dimensional are acquired during the scheduled follow-up visits at 6-month intervals, for example, months = 6, 12, 18, … , 144. However, participants may arbitrarily miss a few of their pre-scheduled visits due to various reasons. As a result, the number of repeated measurements varies by subject, causing an irregular and sparse structure in the data. Such data intrinsic characteristics should be incorporated into the procedure of imbalanced classification, but the increasing difficulty and challenge in implementation is then expected.

To deal with imbalanced classification, one popular approach in the literature is the data-based approach which aims at re-balancing the class distribution by simply resampling the data, such as undersampling the majority class[1] or over-sampling the minority class.[2,3] Nevertheless, this method may either cause a loss of information in the majority class or overuse the data from the minority class. Another popular approach is the algorithm-based approach which mainly depends on the choice of an appropriate inductive bias.[4,5] For instance, different penalties are assigned to different classes in the support vector machine (SVM)-based classifiers.[6] But this type of approach often requires a thorough knowledge of the learning algorithm and the specific application domain, which may be a daunting task to analysts. Another approach is the cost-sensitive approach which considers the varying costs of different misclassification types;[7,8] however these types of costs are usually unknown in practice. Other remedies for imbalanced classification are mainly boosting-based ensemble methods proposed in the area of data science. The boosting algorithms in these methods are basically centered around the combination of several simple classifiers/approaches in order to modify the training data sets for better prediction.[9-14]

To address the classification for high-dimensional data, several approaches have been proposed over the past decade. For example, Fan and Fan[15] proposed the features annealed independence rules (FAIR) to select the most important features via a two-sample *t*-test. Fan and Song[16] established a maximum-marginal-likelihood-type approach for feature screening. Mai and Zou[17] developed the Kolmogorov filter which enjoys the sure screening property to identify statistically significant variables. The fundamental idea of this filter is to construct a specific rule for dimension reduction and use the screened features for subsequent analysis. With application to high-dimensional omics data, Yu and Park[18] proposed an AUC-based approach with penalization such as LASSO and elastic net. Nonetheless, these methods are not capable of dealing with the longitudinal and/or imbalanced structure in data. To handle the classification for longitudinal data, Tomasko et al[19] and Marshall and Barón[20] proposed a modified classical linear discriminant analysis using mixed-effects models to accommodate the over-time underlying associations. De La Cruz-Mesia and Quintana[21] considered a nonlinear hierarchical structure to accommodate the longitudinal profiles and developed a fully Bayesian approach for parameter estimation. More recently, Arribas-Gil et al[22] considered a semiparametric linear mixed-effects model (SLMM) and proposed a unified estimation procedure based on a penalized EM-type algorithm. However, these methods usually require specific distributional assumptions on biomarkers.

These stated methods can only address parts of the issues for complex imbalanced data. To our best knowledge, there is no single approach yet that can accommodate all aforementioned complications comprehensively. As we are motivated by the ADNI study, it is particularly of interest to detect Alzheimer's disease (AD) earlier with all available patient data. Early detection and diagnosis of AD have become increasingly critical for developing future care and treatment. That is because early intervention with medications may slow the progression of disease[23] and provide more opportunities for medical caregivers to gain more understanding about AD and plan for the future. To delay the onset or slow the progression by giving the timely intervention of AD, a prognostic model that can be used for early detection is therefore urgently needed. However, the prevalence of AD in the US elder population (for 65+ year) is 11%,[24] meaning that the class distribution is expected to be skewed and imbalanced. As an evidence, we do observe such a highly skewed distribution in the ADNI data. In the same dataset, we also observe that the brain imaging data which are in high-dimensional and longitudinal setting are collected irregularly and sparsely, which further escalates the challenge of classification as we mentioned previously.

In this article, we propose a two-stage approach to overcome these challenges in classification for complex imbalanced data. Specifically, the techniques of functional principal component analysis (FPCA) are employed for feature extraction from longitudinal biomarkers and then the univariate exponential loss function coupled with group LASSO penalty is used to approximate the empirical area under the receiving operator characteristic curve (ie, AUC) in high-dimensional settings. In other words, the longitudinal data can be first analyzed by FPCA with a significant reduction in its longitudinal dimension, and then the major principal components which are treated as the extracted features can be further used for classification using the proposed AUC-type classifier with group LASSO penalty for feature selection. Finally, the block-wise coordinate descent algorithm is adopted in the process of model estimation. Our approach can substantially

improve the sensitivity that oftentimes is very low for imbalanced data and relieve the computational complexities under such a sophisticated data structure.

For illustration, we apply our approach to ADNI data for early detection of AD. We mainly focus on the participants who are diagnosed as cognitive normal (CN) at baseline but convert to AD at a later time point. To this end, our model is trained to identify the AD patients only using the data right before their first diagnosis of AD. In other words, our approach can early determine high-risk patients who actually have AD later in the near future.

The rest of this article is organized as follows. In Section 2, we briefly introduce the FPCA approach which is often used for dimension reduction in functional data analysis and then present the proposed AUC-type classification framework. In Section 3, we illustrate the proposed classification method using the ADNI data including longitudinal brain imaging data and clinical biomarker data. In Section 4, we conduct extensive simulations to evaluate the performance of the proposed method in finite sample size. Finally, conclusion and possible extensions are discussed in Section 5.

2 | MAIN FRAMEWORK

Our method is a two-stage approach which first involves the use of FPCA to address the longitudinal structure and then uses the proposed AUC-type classifier coupled with group LASSO penalty to improve imbalanced classification. In this section, we briefly introduce the FPCA and empirical AUC, and then present the AUC-type classifier with group LASSO penalty for appropriate variable selection under class imbalance.

2.1 | Functional principal component analysis

To perform a FPCA on irregular and sparse longitudinal data, we adopt a version of FPCA proposed by Yao et al,[25] referred as Principal components Analysis through Conditional Expectation (PACE). Unlike classical FPCA, their approach is particularly useful to model irregular and sparse longitudinal data. The PACE ensures that the functional principal component (FPC) scores extracted from longitudinal features of each subject are well-approximated even when only few measurements are available for a subject. These FPC scores then can be treated as important features/biomarkers summarized from the longitudinal profiles of corresponding subjects[26,27] and used for classification subsequently.

Assume that $M_{ij}(t)$ is the longitudinal trajectory of the jth predictor of the ith subject with $t \in \{1, \dots, T_i\}$. Let $\mu_j(t)$ be its mean function and $\Sigma_j(t,t') = \text{cov}(M_{ij}(t), M_{ij}(t'))$ denote the covariance function which quantifies the correlation between time points t and t'. According to the spectral decomposition, the covariance function can be written as $\Sigma_j(t,t') = \sum_{v=1}^{\infty} \lambda_{jv}\phi_{jv}(t)\phi_{jv}(t')$, where $\{\lambda_{jv}\}_{v=1,\dots,\infty}$ are nonincreasing eigenvalues, that is, $\lambda_{j1} \geq \cdots \geq \lambda_{j\infty} \geq 0$, and $\{\phi_{jv}\}_{v=1,\dots,\infty}$ are the corresponding orthonormal eigenfunctions.

Using the Karhunen-Loève (KL) expansion,[28,29] $M_{ij}(t)$ can be expressed as

$$M_{ij}(t) = \mu_j(t) + \sum_{v=1}^{\infty} \xi_{ijv}\phi_{jv}(t),$$

where $\{\xi_{ijv}\}_{v=1,\dots,\infty}$ are uncorrelated random variables with mean zero and variance λ_{jv}. In practice, $M_{ij}(t)$ is usually approximated by the first $\mathcal{V}$ eigenfunctions as

$$M_{ij}(t) \approx \mu_j(t) + \sum_{v=1}^{\mathcal{V}} \xi_{ijv}\phi_{jv}(t),$$

where $\mathcal{V}$ can be determined by the pre-specified percentage of variance explained (PVE). Specifically, the value of $\mathcal{V}$ is often chosen as the smallest integer such that $\sum_{v=1}^{\mathcal{V}} \lambda_{jv} / \sum_{v=1}^{\infty} \lambda_{jv} \geq \text{PVE}$.

In general, $M_{ij}(t)$ is often observed at irregular and sparse time points. Suppose $U_{ij}(t)$ is a random observation of $M_{ij}(t)$, we have

$$U_{ij}(t) = M_{ij}(t) + \epsilon_{ij}(t),$$

where $\epsilon_{ij}(t)$ is the measurement error with mean zero and variance σ^2. By applying PACE to the jth longitudinal predictor in the pooled data, the estimated mean function $\hat{\mu}_j(t)$, covariance function $\hat{\Sigma}_j(t,t')$, eigenvalues $\hat{\lambda}_{jv}$, eigenfunctions $\hat{\phi}_{jv}(t)$

and error variance $\hat{\sigma}^2$ can be obtained hierarchically. Specifically, $\hat{\mu}_j(t)$ and $\hat{\Sigma}_j(t,t')$ are first estimated using the penalized spline fit and moments approaches as described in the articles of Staniswalis and Lee[30] and Yao.[31] Then $\hat{\lambda}_{jv}$ and $\hat{\phi}_{jv}(t)$ can be obtained from the spectral decomposition of the estimated $\hat{\Sigma}_j(t,t')$. The estimated error variance $\hat{\sigma}^2$ is calculated from the average difference of the middle 60% of diagonal elements between the raw and estimated covariance matrices.[32] Finally, FPC scores $\{\xi_{ijv}\}'$s for the ith subject are estimated as follows:

$$\hat{\xi}_{ijv} = \hat{\lambda}_{jv}\hat{\phi}_{ijv}^T\hat{\Sigma}_{U_{ij}}^{-1}(U_{ij} - \hat{\mu}_{ij}),\ v = 1, 2, \dots, \mathcal{V},$$

where $\hat{\mu}_{ij} = \{\hat{\mu}_j(t)\}_{t=1,\dots,T_i}$ and $\hat{\phi}_{ijv} = \{\hat{\phi}_{jv}(t)\}_{t=1,\dots,T_i}$ are $T_i \times 1$ vectors, and $\hat{\Sigma}_{U_{ij}} = \hat{\Sigma}_j(t,t') + \hat{\sigma}^2\delta_{tt'}$ is a $T_i \times T_i$ matrix with $\delta_{tt'} = 1$ if $t = t'$ and $\delta_{tt'} = 0$ if $t \neq t'$ with $t, t' \in \{1, \dots, T_i\}$. Note that all these FPC scores can be obtained by using the `fpca.sc` function[30,32,33] in the R package `refund`, and $\mathcal{V}$ can be determined by setting a specific value for PVE, such as 90%, 95%, or 99%. Based on what we have observed from the simulations and real data analyses, using $\mathcal{V} = 2$ is generally sufficient enough to characterize the longitudinal data and can simplify the process of extracting features from longitudinal biomarkers using FPCA. With a sensitivity study (not shown here), we notice that the classification performance of our proposed method is not affected by the selection of $\mathcal{V}$, only showing very mild differences in performance. Therefore, we adopt $\mathcal{V} = 2$ for all simulations and real data analysis throughout the article. After obtaining theses FPC scores, a classification procedure can be applied subsequently.

2.2 | Empirical AUC and its surrogate losses

The area under the receiver operating characteristic (ROC) curve, that is, the AUC, is a well-known rank-based statistic and frequently used to evaluate the performance of a classifier. The AUC summarizes both the sensitivity (or true positive rate, TPR) and 1-specificity (or the false positive rate, FPR) and reflects all possible trade-offs between TPR and FPR by varying the decision threshold. Thus, maximizing the AUC is indeed a process of searching for an optimal threshold that leads to both optimal sensitivity and specificity. Because of this, AUC that represents a probability of a randomly selected positive instance having a higher score than a randomly chosen negative instance is thus insensitive to class prevalence and misclassification costs under data imbalance.[34,35]

After extracting FPC scores from the trajectories of all biomarkers, we can combine them linearly, as other traditional AUC-based approaches, to improve prognostic accuracy. The ultimate goal of our study is to find the optimal linear combination of these FPC scores so that the empirical AUC is maximized even under the complex and imbalanced data structure, and hence achieving optimal sensitivity and specificity.

Let X_r^H and X_s^D be a p-dimensional vector containing all FPC scores for the rth and sth subjects in the health and disease groups, respectively, where $r = 1, \dots, n_h$, $s = 1, \dots, n_d$, and n_h and n_d denote the number of subjects in the two groups, respectively. Given any coefficients vector β, the empirical AUC for multiple FPC scores can be estimated as follows:

$$\widehat{AUC}(\beta) = \frac{1}{n_h n_d}\sum_{r=1}^{n_h}\sum_{s=1}^{n_d} I(\beta^T X_r^H < \beta^T X_s^D),$$

where $I(\cdot)$ is the indicator function. However, this estimated empirical AUC can not be used directly for classification in high-dimensional settings because of computational concerns.

Due to the discontinuity and non-convexity of empirical AUC, a widely used technique for circumventing the computational challenge is to approximate the empirical AUC with some pairwise convex surrogate loss function.[36-40] However, it usually necessitates pairwise comparisons between positive and negative instances, resulting in quadratic computational complexity.[41-44] To alleviate the computational burden associated with pairwise surrogate losses, several non-pairwise strongly proper losses, such as the exponential loss and squared hinge loss, have been proposed and shown to be consistent with the AUC maximization task.[42,45,46] Besides that, Gao and Zhou[47] developed a sufficient condition for AUC consistency and established the equivalence of univariate exponential accuracy loss and pairwise exponential surrogate accuracy loss. As a result, using empirical AUC or univariate exponential loss in classification is expected to be equivalent in terms of performance. Thus, we use univariate exponential loss to develop the proposed AUC-type classifier.

2.3 | The proposed AUC-type classification framework

In light of the established equivalence between minimizing the univariate exponential loss and maximizing the empirical AUC, the loss function used in our approach is given as follows to address the issue of class imbalance:

$$l(\beta) = \sum_{i=1}^{N} e^{-y_i x_i^T \beta}, \tag{1}$$

where x_i is a vector containing all FPC scores of ith subject, y_i is the corresponding response with binary outcomes, that is, $y_i = 1$ if positive and $y_i = -1$ if negative,[46] and N denotes the total number of subjects with $N = n_h + n_d$.

Notice that each biomarker trajectory of a subject is summarized as a set of FPC scores. Thus, this set of scores is treated as a grouped feature. Owing to high-dimensional settings, we adopt the group lasso penalty proposed by Yuan[48] to accommodate the grouping structure and perform group-feature selection. The objective function can be written as:

$$l_\tau(\beta) = \frac{1}{N}\sum_{i=1}^{N} e^{-y_i x_i^T \beta} + \tau \sum_{g=1}^{G} \sqrt{p_g}||\beta_g||_2, \tag{2}$$

where β_g is a coefficient vector corresponding to the gth grouped feature, p_g is the number of FPC scores within the gth group, G is the total number of groups, τ is the tuning parameter, and $||\cdot||_2$ is the L_2 norm. Here, $\sqrt{p_g}$ is used to adjust for the varying group sizes. Note that the tuning parameter τ can be determined using a D-fold cross-validation with empirical AUC or univariate exponential loss, which are indeed equivalent in terms of classification performance (see Section 2.2). By the ease of interpretation of AUC, we use empirical AUC as criterion for all simulations and real data analysis throughout the article.

Regarding the choice of D, it generally involves a trade-off between bias and variance. To be more precise, a large value of D typically results in small bias but large variance when evaluating the model performance, whereas a small value of D results in relatively large bias but small variance. The most commonly used values for D are $D = 3$, 5, or 10. Considering the small sample size in the disease group under data imbalance, we adopt a five-fold cross-validation in the following analyses, which not only achieves the bias-variance trade-off but also generates a moderate-sized hold-out fold for validation. In general, one may select a proper D-fold cross-validation based on the sample size and the severity of imbalance.

To solve for the β that minimizes Equation (2), we employ a quadratic approximation which is similar to that in the article of Simon et al.[49] Let $m = X\beta$, where $X = [x_1, x_2, \ldots, x_N]^T$ is the design matrix, and $\dot{l}(\beta)$, $\ddot{l}(\beta)$, $l'(m)$, $l''(m)$ be the gradient and Hessian of the loss function in Equation (1) with respect to β and m, respectively. Using a second-order Taylor expansion centered at the initial value $\tilde{\beta}$, Equation (1) becomes:

$$\begin{aligned} l(\beta) &\approx l(\tilde{\beta}) + (\beta - \tilde{\beta})^T \dot{l}(\tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T \ddot{l}(\beta)(\beta - \tilde{\beta}) \\ &= l(\tilde{\beta}) + (X\beta - \tilde{m})^T l'(m) + \frac{1}{2}(X\beta - \tilde{m})^T l''(m)(X\beta - \tilde{m}) \\ &= \frac{1}{2}(z(\tilde{m}) - X\beta)^T l''(\tilde{m})(z(\tilde{m}) - X\beta) + C(\tilde{m}, \tilde{\beta}), \end{aligned}$$

where $\tilde{m} = X\tilde{\beta}$, $z(\tilde{m}) = \tilde{m} - l''(\tilde{m})^{-1} l'(\tilde{m})$, and $C(\tilde{m}, \tilde{\beta})$ consist of all terms that do not depend on β. Then, $\hat{\beta}$ can be estimated by optimizing a penalized reweighted least squares:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L_\tau(\beta),$$

where

$$L_\tau(\beta) = \frac{1}{2N}(z(\tilde{m}) - X\beta)^T l''(\tilde{m})(z(\tilde{m}) - X\beta) + \tau \sum_{g=1}^{G} \sqrt{p_g}||\beta_g||_2.$$

The objective function $L_\tau(\beta)$ consists of a quadratic term and the group lasso penalty. The quadratic term can be viewed as squared errors in the estimated $\hat{\beta}$ between the current and previous iterations. As we aim to minimize $L_\tau(\beta)$, the estimator $\hat{\beta}$ is viewed as an solution with the least squared error to maximize the empirical AUC. Regarding the term of group

lasso penalty, it intrinsically ensures that only a subset of "group" features are selected, thus significantly reducing the model complexity. Each of $\{\beta_g\}_{g=1,\dots,G}$ can be estimated iteratively by the block coordinate descent algorithm presented by Yuan.[48] Specifically, to solve for the coefficients vector β_k for the kth grouped feature, we first compute the corresponding first derivative of $L_\tau(\beta)$ as:

$$\frac{\partial L_\tau(\beta)}{\partial \beta_k} = -\frac{1}{N} X_k^T l''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq k} X_g \beta_g - X_k \beta_k \right) + \tau \sqrt{p_k} s_k, \tag{3}$$

where X_g and X_k are the data matrices corresponding to the gth and kth grouped features respectively, p_k is the group size of kth grouped feature, and

$$\begin{cases} s_k = \frac{\beta_k}{||\beta_k||_2}, & \text{if } \beta_k \neq \mathbf{0} \\ ||s_k||_2 \leqslant 1, & \text{if } \beta_k = \mathbf{0}. \end{cases}$$

Next, by setting Equation (3) to zero, we can obtain $\hat{\beta}_k$. Specifically, when $\beta_k = \mathbf{0}$, we can get:

$$\left\| \frac{1}{N} X_k^T l''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq k} X_g \beta_g \right) \right\|_2 \leqslant \tau \sqrt{p_k}, \tag{4}$$

when $\beta_k \neq \mathbf{0}$, it is easy to obtain:

$$\hat{\beta}_k = \left[\frac{1}{N} X_k^T l''(\tilde{m}) X_k + \frac{\tau \sqrt{p_k}}{||\beta_k||_2} \cdot I \right]^{-1} \cdot \left[\frac{1}{N} X_k^T l''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq k} X_g \beta_g \right) \right]. \tag{5}$$

Hence, cycling through each group of FPC scores, simultaneous variable selection, and model estimation can be achieved via the following Algorithm 1.

Algorithm 1.

Step 1. Initialize $\tilde{\beta}$, and compute $\tilde{m}$, $l'(\tilde{m})$, $l''(\tilde{m})$, and $z(\tilde{m})$.
Step 2. For $k = 1, \dots, G$, if Equation (4) holds, $\hat{\beta}_k$ is set to $\mathbf{0}$; otherwise, $\hat{\beta}_k$ is updated using Equation (5).
Step 3. Set $\tilde{\beta} = \hat{\beta}$, and compute $\tilde{m}$, $l'(\tilde{m})$, $l''(\tilde{m})$, and $z(\tilde{m})$.
Step 4. Repeat Steps 2-3 until convergence.

It is worth mentioning that the proposed objective function is guaranteed to converge to the global minimum using the above algorithm when initialized with an arbitrary value for $\tilde{\beta}$. The detailed convergence analysis has been thoroughly discussed by Tseng.[50] To reduce the number of required iterations and increase the computational efficiency in high-dimensional sparse settings, we suggest initializing $\tilde{\beta}$ with a vector of small values, such as $\tilde{\beta} = (0.001, \dots, 0.001)$ as we used in this article.

To regularize with the group lasso penalty, variable selection is conducted on the group level. Specifically, each set of FPC scores simply represents each longitudinal biomarker. Therefore, these scores extracted from a particular biomarker can be only all selected or all dropped, depending on whether the associated biomarker is important or not to the model. To speed up the computation, we employ a strategy called *active-set* convergence which has been discussed in the articles of Krishnapuram et al,[51] Meier et al,[52] and Friedman et al.[53] Specifically, after the first cycle through G groups, the remaining iterations will be restricted to the *active-set* which will be updated after each cycle. The entire process stops after the *active-set* does not change.

3 | ALZHEIMER'S DISEASE DATA

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal

Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Detailed information regarding the ADNI study and the complete protocol can be found in the articles of Mueller et al[54] and Jack et al.[55] In the ADNI data, participants are labeled with: cognitively normal (CN), MCI, or AD based on a series of assessments at their initial visits. These are also their states at baseline. It is expected to have repeated evaluations conducted subsequently at a 6-month interval.

Most existing studies focused on predicting the conversion from MCI to AD for individuals who were diagnosed as MCI at baseline. However, the conversion process could begin years before the onset of symptoms. In our analysis, we focus on the development of a prognostic model that can be used for early detection of AD among CN individuals. We select 267 subjects who are normal at baseline and have at least three visits. Among them, 30 subjects progress to AD at a later time, denoted as AD, and 237 subjects remain as normal, denoted as CN. The demographic information of those subjects is summarized in Table 1. It should be noted that the longitudinal data are indeed irregularly observed among participants. Specifically, each participant undergoes these assessments at different time points and has a different number of visits. The distribution of number of visits is presented in Table 2.

In the literature, biomarkers from different modalities have been utilized to investigate the progression of AD. Brain abnormalities detected by MRI are considered to be valid markers of AD and are widely used to predict the conversion from MCI to AD.[56-60] Fluorodeoxyglucose positron emission tomography (FDG-PET) is able to provide the estimates of cerebral metabolic rates of glucose, thus revealing the pattern of regional hypometabolism which is a prominent hallmark of AD.[61-64] Additionally, biomedical changes in the brain are directly presented in the Cerebrospinal fluid (CSF). Hence, CSF-based biomarkers are often employed to depict the pathological changes of AD.[65-68]

In this study, we mainly focus on biomarkers that are extracted from the MRI modality. All of the 3D T1-weighted MRI images downloaded from the ADNI database for each subject are processed using Freesurfer (v6.0.0, Martinos Center for Biomedical Imaging) which is an open-source software suite and freely available at *FreeSurferWiki* (https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki). The longitudinal processing pipeline in Freesurfer consists of the following steps: spatial normalization and intensity correction, Talairach registration, brain mask creation, subcortical segmentation, surfaces reconstruction, and cortical atlas registration and parcellations. More details about the processing framework can be found in the article of Reuter.[69] There are 319 biomarkers in total generated by Freesurfer v6.0.0,

TABLE 1 Demographic characteristics of selected subjects

		Age (years)		Gender (%)	
Group	**n**	**Mean**	**SD**	**Male**	**Female**
CN	237	74.5	5.6	52.7	47.3
AD	30	75.4	3.9	40.0	60.0

TABLE 2 Distribution of number of visits

	Number of subjects	
Visits	**CN**	**AD**
3	68	2
4	100	4
5	13	3
6	10	5
7	13	2
8	14	4
9	10	7
10	9	3
Total	237	30

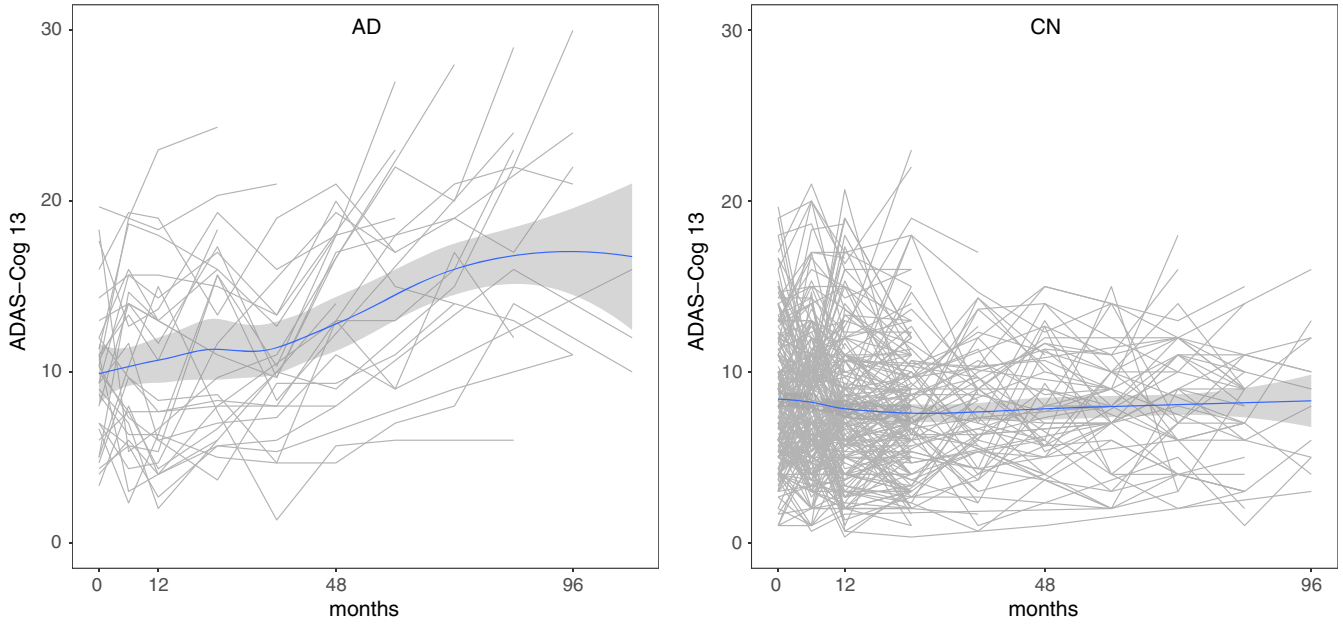

FIGURE 1 Longitudinal trajectories of ADAS-Cog 13 for cognitively normal subjects and AD patients

with each corresponding to a specific region of interest (ROI) in the brain. More specifically, these ROIs consist of cortical volume, cortical thickness average, cortical surface area, and the volume estimates of a wide range of subcortical structures.[70,71]

In addition to those biomarkers extracted from the brain imaging data, we also include five cognitive and functional scores which are closely associated with AD and popular in the literature:[27,72,73] Alzheimer's Disease Assessment Scale-Cognitive 13 items (ADAS-Cog 13), Mini Mental State Examination (MMSE), Functional Assessment Questionnaire (FAQ), Rey Auditory Verbal Learning Tests (RAVLT immediate score and RAVLT learning score). Besides that, other demographic and genetic variables that might be predictive of AD conversion are also included: baseline age, gender, and apolipoprotein E allele $\varepsilon 4$ (APOE4). Figure 1 presents the longitudinal trajectories of ADAS-Cog 13 for subjects used in this study, showing the sparse and irregular characteristics of the ADNI dataset. The trends in the two plots suggest the potential of using ADAS-Cog 13 to identify AD patients among these normal subjects at baseline.

For the model training, the last visit data of each CN subject is excluded. But for AD patients, we use the data before their first diagnosis of AD in order to train the model only based on the data before progressing to AD. By this, our model is capable of identifying potential AD patients before their next clinical visit. As an illustration, the data of a CN (or an AD) participant that is used for model training is shown in Figure 2 with a red box.

For the model evaluation, the processed data are randomly split into training and test subsets, comprising 70% and 30% of all instances respectively. A stratified sampling method is employed to ensure that both subsets have the same imbalance ratio.[74] To deal with these longitudinal biomarkers, the PACE algorithm proposed by Yao[25] is applied to obtain the corresponding FPC scores which are then used as predictors in our model. The tuning parameter in the proposed method is selected by five-fold cross-validation using the empirical AUC as the criterion. For comparison purposes, logistic regression with L_1 penalty and SVM with linear kernel are also conducted with this ADNI dataset. The results based on 500 Monte Carlo replicates are given in Table 3. It is worth noting that the class distribution is highly imbalanced in this ADNI dataset (ie, CN=237, AD=30). Both penalized logistic regression and SVM are biased towards the majority class, thus leading to the low sensitivity of 36% and 44%, respectively. Moreover, it seems that SVM tends to overfit under the high-dimensional setting and performs poorly on the test data. However, our proposed approach is capable of dealing with the case of class imbalance, and achieves superior classification performance, especially in terms of sensitivity which is often considered as an important measure in medical diagnosis. As shown in Table 3, the performance of the proposed framework outperforms L_1 logistic regression and linear SVM in terms of its AUC and sensitivity (88% and 79%, respectively) with a slight compromise in specificity, which indicates the superiority of our method for such a complex imbalanced dataset. Finally, our approach indicates that several biomarkers selected by group LASSO seem associated with early detection of AD. For example, the biomarkers with high absolute value of coefficient include: FAQ and ADAS in clinical scores; left and right postcentral gyrus, left precentral gyrus in subcortical volumes; left postcentral gyrus,

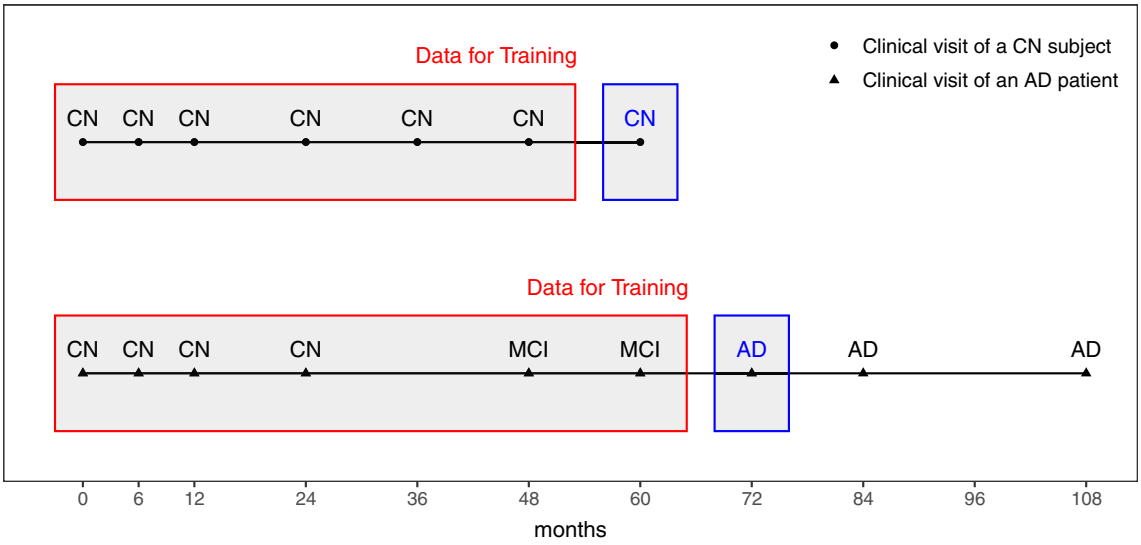

FIGURE 2 Clinical diagnosis of a CN subject or an AD patient over time. The red box represents the data used for model training. The blue box represents the final diagnosis used as the membership outcome

TABLE 3 Classification results (S.E.) for ADNI data with L_1 logistic, linear SVM and the proposed method based on 500 Monte Carlo replicates

		L_1 logistic	Linear SVM	Proposed method
Training set	Sensitivity	0.601 (0.297)	0.999 (0.001)	0.946 (0.066)
(n_h=166, n_d=21)	Specificity	0.999 (0.001)	0.999 (0.001)	0.973 (0.035)
	Accuracy	0.956 (0.033)	0.999 (0.001)	0.970 (0.035)
	AUC	0.918 (0.167)	0.999 (0.001)	0.976 (0.033)
Test set	Sensitivity	0.362 (0.199)	0.441 (0.154)	0.790 (0.145)
(n_h=71, n_d=9)	Specificity	0.996 (0.008)	0.980 (0.015)	0.880 (0.094)
	Accuracy	0.925 (0.022)	0.919 (0.023)	0.870 (0.084)
	AUC	0.832 (0.147)	0.854 (0.068)	0.880 (0.091)

Abbreviations: L_1 logistic, logistic regression with L_1 penalty; Linear SVM, support vector machine with linear kernel; (n_h, n_d), number of subjects in the CN and AD groups respectively.

right medial orbitofrontal cortex, right supramarginal gyrus, right pericalcarine cortex in cortical thicknesses. Albeit interesting, more thorough investigations from the view of neuroscience are strongly encouraged before coming to any further conclusions.

4 | SIMULATION STUDY

In this section, we conduct extensive simulations to evaluate the performance of the proposed method. Two data-generating schemes are considered: (1) class memberships are generated by a logistic regression model; (2) class memberships are pre-determined by the belonging group: health or disease. For each scheme, the classification performance is further assessed under two settings: (i) a low-dimensional setting with $n > p$ and (ii) a high-dimensional setting with $n < p$.

Throughout all simulations, it is assumed that each subject has a longitudinal profile with observations measured at seven different time points (ie, $t \in \{0, 1, 2, 3, 4, 5, 6\}$ and $t = 0$ represents the baseline). We also perform other two popular methods (ie, logistic regression and SVM) at various levels of class imbalance for comparison purposes.

4.1 | Class memberships determined by a logistic regression model

In the first scheme, we generate class memberships using a logistic regression model. Specifically, it is a two-stage process. In the first stage, we assume that the jth longitudinal predictor $U_{ij}(t)$ for the ith subject is generated by a linear model:

$$U_{ij}(t) = \gamma_{0j} + \gamma_{1j}t + \gamma_{2j}t^2 + b_{ij} + \epsilon_{ij}(t),\ t \in \{0, 1, 2, \ldots, 6\},$$

where the subject-specific random effect b_{ij} is generated from $N(0, 1.3)$ and the measurement error ϵ_{ij} is generated from $N(0, 1)$. In the second stage, we convert the longitudinal predictor $U_{ij}(t)$ into a set of FPC scores using the FPCA approach, then denoted as x_{ij}. These sets of FPC scores are indeed considered as features and then used in the subsequent classification procedure. As we extract the FPC scores using the PACE algorithm, the number of principal components is fixed at two, for simplicity, to override the required setting for PVE.

Later, the class memberships are assigned through the following logistic regression model:

$$Y_i = \begin{cases} 1, & \text{if } \left[1 + \exp\left\{-\beta_0 - \sum_j x_{ij}^T \beta_j\right\}\right]^{-1} > 0.5 \\ 0, & \text{if } \left[1 + \exp\left\{-\beta_0 - \sum_j x_{ij}^T \beta_j\right\}\right]^{-1} \leq 0.5, \end{cases}$$

where β_j is the coefficient vector corresponding to the jth longitudinal predictor. Notice that β_j is a 2×1 vector as we set the number of principal components at two for each longitudinal predictor. The intercept β_0 can be set to generate different levels of class imbalance. Typically, the membership can be coded as "health" if $Y_i = 0$ and "disease" if $Y_i = 1$.

For our analysis, low and high dimensional settings are examined separately. For each setting, 500 Monte Carlo replicates are simulated at each imbalance ratio. For each replicate, the data of 600 subjects are generated. Among them, 300 subjects are used for model training and the rest of 300 are used as a test data set for evaluation.

(i) *Low-dimensional setting:* Three (3) longitudinal predictors are simulated for each subject, where we set $(\gamma_{01}, \gamma_{11}, \gamma_{21})^T = (1.5, -0.25, 0.1)^T$, $(\gamma_{02}, \gamma_{12}, \gamma_{22})^T = (1, -0.2, 0.11)^T$, and $(\gamma_{03}, \gamma_{13}, \gamma_{23})^T = (2, -0.15, 0.09)^T$. To obtain class memberships using the above logistic regression model, we let $\beta_1 = (-2, 1)$, $\beta_2 = (-1, 0.5)$, $\beta_3 = (1.5, -1)$. The intercept β_0 is given by different values ($\{-2.5, -3.5, -4.5\}$) to obtain the imbalance ratio of $\{3.2, 5.3, 9.0\}$, respectively. The classification results are presented in Table 4. In this setting, the performances of three methods are comparable in terms of AUC and accuracy. However, regardless of training or testing, noticeable lower sensitivities are observed in the methods of logistic regression and SVM as the imbalance ratio increases, whereas the sensitivity declines slightly with the proposed method.

(ii) *High-dimensional setting:* Five hundred (500) longitudinal predictors are simulated for each subject, where the coefficients of $\{\gamma_{qj}\}_{q=0,1,2}$ that correspond to the jth predictor are generated randomly from truncated normal distributions (TN):

$$\begin{aligned} \gamma_{0j} &\sim TN(1.5, 1),\ \gamma_{0j} \in [1, 2], \\ \gamma_{1j} &\sim TN(-0.15, 1),\ \gamma_{1j} \in [-0.2, -0.1], \\ \gamma_{2j} &\sim TN(0.11, 1),\ \gamma_{2j} \in [0.09, 0.13]. \end{aligned}$$

For simplicity, we assume that the first five predictors are significant, with each corresponding β_j specified as follows: $\beta_1 = (1.5, -0.5)$, $\beta_2 = (-1.2, -1.5)$, $\beta_3 = (-0.5, 1)$, $\beta_4 = (0.5, -1)$, $\beta_5 = (-1.5, 1)$. The remaining 495 predictors are assumed to be insignificant, thus having $\beta_j = (0, 0)$, $j \in \{6, \ldots, 500\}$. Similar to the low-dimensional setting above, different levels of class imbalance (imbalance ratio = $\{3.2, 4.9, 6.1\}$) are assessed by assigning different values ($\{-3, -4, -4.5\}$) for β_0 correspondingly. The simulation results are provided in Table 5. In this setting, the performance of the proposed method is better than that of the other two approaches in terms of AUC and sensitivity. It seems that logistic regression and SVM tend to classify subjects into the majority class (ie, the health group), thus resulting in low sensitivity. However, the proposed method achieves a better sensitivity with a little sacrifice of specificity and accuracy.

TABLE 4 Classification results (S.E.) of L_1 logistic regression, linear SVM and the proposed method at various imbalance ratios in low-dimensional setting based on 500 Monte Carlo replicates

		$n_h/n_d = 3.2$			$n_h/n_d = 4.9$			$n_h/n_d = 6.1$		
Imbalance ratio		**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**
Training	Sensitivity	0.689	0.681	0.873	0.618	0.594	0.884	0.548	0.483	0.896
		(0.061)	(0.071)	(0.035)	(0.081)	(0.100)	(0.039)	(0.108)	(0.156)	(0.041)
	Specificity	0.941	0.945	0.856	0.965	0.970	0.867	0.981	0.987	0.882
		(0.013)	(0.015)	(0.034)	(0.009)	(0.012)	(0.034)	(0.007)	(0.008)	(0.038)
	Accuracy	0.880	0.882	0.860	0.910	0.910	0.870	0.938	0.937	0.883
		(0.019)	(0.019)	(0.025)	(0.016)	(0.016)	(0.029)	(0.013)	(0.013)	(0.034)
	AUC	0.933	0.932	0.932	0.940	0.937	0.937	0.948	0.945	0.945
		(0.016)	(0.016)	(0.016)	(0.017)	(0.018)	(0.017)	(0.017)	(0.019)	(0.018)
Test	Sensitivity	0.672	0.658	0.833	0.594	0.564	0.828	0.507	0.435	0.819
		(0.065)	(0.071)	(0.059)	(0.087)	(0.097)	(0.074)	(0.111)	(0.147)	(0.087)
	Specificity	0.933	0.935	0.840	0.959	0.964	0.857	0.975	0.981	0.871
		(0.021)	(0.021)	(0.039)	(0.015)	(0.016)	(0.037)	(0.012)	(0.013)	(0.040)
	Accuracy	0.870	0.869	0.838	0.900	0.899	0.852	0.927	0.925	0.866
		(0.018)	(0.018)	(0.026)	(0.016)	(0.016)	(0.028)	(0.015)	(0.015)	(0.032)
	AUC	0.923	0.922	0.921	0.929	0.927	0.927	0.934	0.932	0.931
		(0.017)	(0.018)	(0.018)	(0.019)	(0.019)	(0.020)	(0.020)	(0.020)	(0.021)

Note: $n_h + n_d = 300$.
Abbreviation: n_h, n_d, number of subjects in health and disease groups respectively.

4.2 | Class memberships pre-determined by health and disease groups

Unlike the previous data-generating scheme, we generate class memberships without using any model-based mechanisms. The longitudinal predictors are simulated for the health (H) and disease (D) groups separately:

$$U_{rj}^H(t) = \mu_j^H(t) + b_{rj} + \varepsilon_{rj}(t),\ t \in \{0, 1, \ldots, 6\},$$
$$U_{sj}^D(t) = \mu_j^D(t) + b_{sj} + \varepsilon_{sj}(t),\ t \in \{0, 1, \ldots, 6\},$$

where $\mu_j^H(t)$ and $\mu_j^D(t)$ are the mean functions of the jth longitudinal predictor for the rth health and the sth disease subject, respectively, b_{rj} and b_{sj} are the subject-specific random effects generated from $N(0, 1.5)$. The random errors $\varepsilon_{rj}(t)$ and $\varepsilon_{sj}(t)$ are generated from $N(0, 1)$. The PACE algorithm is applied to each predictor to extract the FPC scores which are further used as features in the proposed method. By this data-generating scheme, class memberships of all subjects are pre-determined, that is, $Y = 0$ if health and $Y = 1$ if disease.

Under this scheme, we also consider low and high-dimensional settings. The classification performances are also examined at different levels of class imbalance. Assuming a total sample size of 300, different numbers of subjects are assigned to the health and disease groups to generate various imbalance ratios. That is, $(n_h, n_d) = \{(225, 75),\ (257, 43),\ (270, 30)\}$ for the ratios of $n_h/n_d = \{3,\ 5.98,\ 9\}$. In each scenario, 500 Monte Carlo replicates are simulated.

(i) *Low-dimensional setting:* Three (3) longitudinal predictors are simulated for each of the subjects. For the health group, the mean μ_j^H is assumed to be constant across different time points. Specifically, we set: $\mu_1^H = (1, 1, 1, 1, 1, 1, 1)^T$, $\mu_2^H = (2, 2, 2, 2, 2, 2, 2)^T$, $\mu_3^H = (1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5)^T$. For the disease group, we let $\mu_j^D = \left\{\gamma_{0j} + \gamma_{1j}t + \gamma_{2j}t^2\right\}$, where $t = 0, 1, \ldots, 6$, to reflect the progression of the disease. Three sets of $\{\gamma_{qj}\}_{q=0,1,2}$ used for

TABLE 5 Classification results (S.E.) of logistic regression, linear SVM and the proposed method at various imbalance ratios in high-dimensional setting based on 500 Monte Carlo replicates

		$n_h/n_d = 3.2$			$n_h/n_d = 4.9$			$n_h/n_d = 6.1$		
Imbalance ratio		**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**
Training	Sensitivity	0.705	0.999	0.900	0.659	0.999	0.905	0.549	0.999	0.899
		(0.221)	(0.001)	(0.035)	(0.293)	(0.001)	(0.052)	(0.360)	(0.001)	(0.062)
	Specificity	0.997	0.999	0.894	0.999	0.999	0.891	0.999	0.999	0.888
		(0.005)	(0.001)	(0.037)	(0.002)	(0.001)	(0.050)	(0.001)	(0.001)	(0.061)
	Accuracy	0.928	0.999	0.896	0.942	0.999	0.893	0.938	0.999	0.889
		(0.053)	(0.001)	(0.032)	(0.049)	(0.001)	(0.047)	(0.049)	(0.001)	(0.058)
	AUC	0.982	0.999	0.957	0.982	0.999	0.952	0.944	0.999	0.946
		(0.022)	(0.001)	(0.020)	(0.051)	(0.001)	(0.034)	(0.135)	(0.001)	(0.044)
Test	Sensitivity	0.412	0.221	0.791	0.262	0.109	0.724	0.174	0.063	0.686
		(0.112)	(0.058)	(0.078)	(0.122)	(0.056)	(0.122)	(0.125)	(0.043)	(0.137)
	Specificity	0.968	0.901	0.856	0.982	0.958	0.860	0.989	0.977	0.860
		(0.021)	(0.026)	(0.039)	(0.016)	(0.017)	(0.048)	(0.013)	(0.013)	(0.057)
	Accuracy	0.836	0.740	0.841	0.859	0.813	0.837	0.874	0.848	0.835
		(0.025)	(0.023)	(0.031)	(0.021)	(0.020)	(0.037)	(0.019)	(0.018)	(0.044)
	AUC	0.892	0.645	0.913	0.876	0.640	0.889	0.842	0.635	0.877
		(0.029)	(0.038)	(0.028)	(0.050)	(0.044)	(0.043)	(0.108)	(0.050)	(0.047)

Note: $n_h + n_d = 300$.
Abbreviation: n_h, n_d: number of subjects in the health and disease groups respectively.

the data generation are specified as follows: $(\gamma_{01}, \gamma_{11}, \gamma_{21})^T = (1, -0.2, 0.08)^T$, $(\gamma_{02}, \gamma_{12}, \gamma_{22})^T = (2, -0.25, 0.07)^T$, and $(\gamma_{03}, \gamma_{13}, \gamma_{23})^T = (1.5, -0.15, 0.09)^T$.

(ii) *High-dimensional setting:* Five hundred (500) longitudinal predictors are simulated for each subject. Among them, the last 475 predictors are considered insignificant and the corresponding mean functions are assumed to be the same for both the health and disease groups, that is, $\mu_j^H = \mu_j^D = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)^T$, $j \in \{26, \ldots, 500\}$. For the first 25 predictors that are considered significant, their mean functions are generated differently for health and disease groups. For the health group, the mean μ_j^H is assumed to be constant, that is, $\mu_j^H = (c_j^H, c_j^H, \ldots, c_j^H)_{1\times 7}^T$, $j \in \{1, \ldots, 25\}$, where c_j^H is generated from a truncated normal distribution (TN):

$$c_j^H \sim TN(0,1),\ c_j^H \in [-1,1].$$

For the disease group, we let $\mu_j^D = \left\{\gamma_{0j} + \gamma_{1j}t + \gamma_{2j}t^2\right\}_{t=0,1,\ldots,6}$, where the coefficients $\{\gamma_{qj}\}_{q=0,1,2}$ that correspond to the jth predictor are randomly selected, for each Monte Carlo sample, from truncated normal distributions:

$$\begin{aligned}
\gamma_{0j} &\sim TN(0,1),\ \gamma_{0j} \in [-1,1],\\
\gamma_{1j} &\sim TN(0,1),\ \gamma_{1j} \in [-0.1,0.1],\\
\gamma_{2j} &\sim TN(0,1),\ \gamma_{2j} \in [-0.01,0.01].
\end{aligned}$$

The simulation results are given in Tables 6 and 7. Even under this data-generating mechanism, the proposed approach outperforms logistic regression and SVM across various levels of class imbalance in many perspectives, especially the good performance in sensitivity regardless of being in low- or high-dimensional setting (see Tables 6 and 7). When the class imbalance becomes more severe, the proposed method still can achieve a high sensitivity whereas a substantial drop is observed in the other two methods. It is worth mentioning that the AUCs of the proposed method are higher than those

TABLE 6 Classification results (S.E.) with various sample sizes of health and disease groups in low-dimensional setting based on 500 Monte Carlo replicates

		$(n_h, n_d) = (225, 75)$			$(n_h, n_d) = (257, 43)$			$(n_h, n_d) = (270, 30)$		
Sample size		**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**
Training	Sensitivity	0.774	0.767	0.895	0.596	0.555	0.873	0.454	0.379	0.852
		(0.097)	(0.108)	(0.049)	(0.205)	(0.250)	(0.071)	(0.226)	(0.278)	(0.086)
	Specificity	0.954	0.957	0.894	0.976	0.982	0.864	0.985	0.991	0.846
		(0.014)	(0.015)	(0.045)	(0.008)	(0.010)	(0.074)	(0.007)	(0.008)	(0.080)
	Accuracy	0.909	0.909	0.895	0.922	0.921	0.866	0.932	0.930	0.846
		(0.032)	(0.032)	(0.042)	(0.029)	(0.032)	(0.070)	(0.021)	(0.023)	(0.079)
	AUC	0.953	0.952	0.951	0.927	0.919	0.925	0.906	0.885	0.902
		(0.033)	(0.035)	(0.034)	(0.064)	(0.081)	(0.064)	(0.076)	(0.110)	(0.077)
Test	Sensitivity	0.751	0.743	0.865	0.555	0.511	0.818	0.416	0.335	0.780
		(0.099)	(0.108)	(0.059)	(0.198)	(0.235)	(0.095)	(0.214)	(0.254)	(0.119)
	Specificity	0.949	0.951	0.882	0.970	0.976	0.853	0.981	0.988	0.837
		(0.017)	(0.016)	(0.048)	(0.012)	(0.014)	(0.075)	(0.010)	(0.011)	(0.080)
	Accuracy	0.899	0.898	0.878	0.911	0.909	0.848	0.925	0.922	0.832
		(0.033)	(0.032)	(0.044)	(0.029)	(0.030)	(0.071)	(0.019)	(0.020)	(0.077)
	AUC	0.945	0.944	0.945	0.912	0.907	0.912	0.889	0.872	0.889
		(0.037)	(0.038)	(0.037)	(0.066)	(0.079)	(0.068)	(0.079)	(0.109)	(0.080)

Note: $n_h + n_d = 300$.
Abbreviation: (n_h, n_d): number of subjects in the health and disease groups respectively.

of logistic regression and SVM in the high-dimensional setting, also coming along with smaller SEs. This result indeed indicates the stability of our approach in high-dimensional settings.

5 | DISCUSSION

In this work, we have developed a novel classification framework for imbalanced data under a longitudinal and high-dimensional structure. With the use of FPCA, a substantial dimension reduction has been achieved for the irregular and sparse longitudinal data, and no distributional assumptions on biomarkers are needed. Unlike other traditional classification methods, the proposed AUC-type classifier with univariate exponential loss function can well and efficiently approximate the empirical AUC which is intrinsically robust against imbalance, thus resulting in a great sensitivity without largely impairing the overall accuracy and specificity. Coupled with the group lasso penalty, feature selection can be conducted within the procedure of classification simultaneously.

As early detection of AD is a recognized health care priority in the United States,[75] we can initially respond to this task by applying the proposed method using the longitudinal brain imaging data together with clinical and cognitive measures. To the best of our knowledge, this is the first study in the literature that focuses on using the longitudinal MRI data to early identify AD patients among these individuals who are diagnosed as normal at baseline. The proposed method not only can detect the at-risk AD patients among these baseline normal-cognition participants but also can identify the most significant biomarkers (such as brain regions) that are associated with the development of AD, though biomarker discovery often requires further and deeper investigations. The proposed method can handle longitudinal and high-dimensional imaging data; however, in practice, each individual's imaging data may not always be available. Because an MRI scan typically is a more expensive procedure which may keep normal individuals from doing the scan and further resulting in the lack of imaging data. But even without the brain imaging data, the proposed method still can perform nicely as we have shown in the simulation study under low-dimensional settings. Apart from the longitudinal data, the

TABLE 7 Classification results (S.E.) with various sample sizes of health and disease groups in high-dimensional setting based on 500 Monte Carlo replicates

		$(n_h, n_d) = (225, 75)$			$(n_h, n_d) = (257, 43)$			$(n_h, n_d) = (270, 30)$		
Sample size		**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**	**Logistic**	**SVM**	**Proposed**
Training	Sensitivity	0.825	0.999	0.927	0.782	0.999	0.921	0.395	0.999	0.911
		(0.121)	(0.001)	(0.033)	(0.234)	(0.001)	(0.039)	(0.369)	(0.001)	(0.053)
	Specificity	0.992	0.999	0.924	0.999	0.999	0.918	0.999	0.999	0.904
		(0.007)	(0.001)	(0.032)	(0.001)	(0.001)	(0.035)	(0.001)	(0.001)	(0.053)
	Accuracy	0.950	0.999	0.924	0.968	0.999	0.918	0.939	0.999	0.905
		(0.034)	(0.001)	(0.028)	(0.034)	(0.001)	(0.032)	(0.036)	(0.001)	(0.051)
	AUC	0.987	0.999	0.972	0.993	0.999	0.968	0.842	0.999	0.956
		(0.013)	(0.001)	(0.015)	(0.024)	(0.001)	(0.019)	(0.224)	(0.001)	(0.034)
Test	Sensitivity	0.657	0.457	0.861	0.389	0.192	0.793	0.131	0.064	0.709
		(0.063)	(0.056)	(0.053)	(0.113)	(0.056)	(0.076)	(0.121)	(0.047)	(0.113)
	Specificity	0.967	0.930	0.892	0.987	0.981	0.895	0.997	0.994	0.885
		(0.015)	(0.016)	(0.031)	(0.009)	(0.008)	(0.037)	(0.004)	(0.005)	(0.054)
	Accuracy	0.890	0.812	0.884	0.901	0.868	0.881	0.910	0.901	0.868
		(0.018)	(0.020)	(0.024)	(0.015)	(0.011)	(0.031)	(0.010)	(0.006)	(0.046)
	AUC	0.947	0.832	0.951	0.922	0.807	0.931	0.777	0.762	0.897
		(0.016)	(0.029)	(0.017)	(0.032)	(0.037)	(0.026)	(0.183)	(0.049)	(0.042)

Note: $n_h + n_d = 300$.
Abbreviation: (n_h, n_d): number of subjects in the health and disease groups respectively.

proposed method can still be applied to cross-sectional imbalanced data, for example applications to the gene expression microarray data,[76] by simply skipping the use of FPCA.

The proposed method is mainly developed for imbalanced classification in longitudinal and high-dimensional settings, but the feature extraction process via FPCA could be somewhat time-consuming when the longitudinal data is dense or the total number of subjects is large. This can be improved by employing other techniques of functional data analysis, for example, the natural cubic spline which has been proven to be an easy-implemented and efficient approach for both sparse longitudinal data and dense functional data.[77-79] Besides that, the FPCA requires a pre-specified number of basis functions, which might be critical for extracting the FPC scores. A simulation was conducted to study how to determine the number of basis functions for FPCA and how the number of basis functions impacts the imbalance classification (results not shown). We suggest using the minimal number of measurements among all subjects minus one as the number of basis functions to ensure the FPC scores can be successfully obtained. It is also worth noting that the feature extraction (ie, FPC scores) by PACE can still be performed even when missing values occur in the longitudinal profiles of subjects. However, because the proposed method is supervised, the response/label must be provided for model training as required in our proposed loss function.

Finally, it is possible to extend our approach to incorporating other alternative surrogate loss functions, such as square and squared hinge losses, for the approximation of the empirical AUC. Such an extension may potentially improve the classification performance and reduce the computational burden. Furthermore, the extension to data that are generated from nonlinear spaces can make the proposed method more general. As one possible solution, a kernelized transformation may be performed on the data prior to any statistical or machine learning modeling. These extensions are indeed beyond the scope of this article and require further investigations.

ACKNOWLEDGEMENTS

We are very grateful for the associate editor and two reviewers' insightful and constructive comments. 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 (http://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.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://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.

ORCID

Yiming Li https://orcid.org/0000-0003-1641-098X
Wei-Wen Hsu https://orcid.org/0000-0002-7822-0826

REFERENCES

1. Japkowicz N. The class imbalance problem: significance and strategies. Paper presented at: Proceedings of the International Conference on Artificial Intelligence (ICAI); Vol. 56, 2000:111-117; Citeseer, Las Vegas, NV, USA.
2. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor Newslett*. 2004;6(1):20-29.
3. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
4. Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recogn*. 2007;40(12):3358-3378.
5. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. *Int J Pattern Recognit Artif Intell*. 2009;23(04):687-719.
6. Lin Y, Lee Y, Wahba G. Support vector machines for classification in nonstandard situations. *Mach Learn*. 2002;46(1):191-202.
7. Margineantu DD. Class probability estimation and cost-sensitive classification decisions. Paper presented at: Proceedings of the European Conference on Machine Learning; 2002:270-281; Springer, New York, NY.
8. Zadrozny B, Langford J, Abe N. Cost-sensitive learning by cost-proportionate example weighting. Paper presented at: Proceedings of the 3rd IEEE International Conference on Data Mining; 2003:435-442; IEEE, Melbourne, FL, USA.
9. Wang BX, Japkowicz N. Boosting support vector machines for imbalanced data sets. *Knowl Inf Syst*. 2010;25(1):1-20.
10. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans Syst Man Cybern A Syst Humans*. 2009;40(1):185-197.
11. Galar M, Fernández A, Barrenechea E, Herrera F. EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recogn*. 2013;46(12):3460-3471.
12. Díez-Pastor JF, Rodríguez JJ, Garcia-Osorio C, Kuncheva LI. Random balance: ensembles of variable priors classifiers for imbalanced data. *Knowl-Based Syst*. 2015;85:96-111.
13. Wan X, Ma P, Zhang X. A promising choice in hypertension treatment: fixed-dose combinations. *Asian J Pharm Sci*. 2014;9(1):1-7.
14. Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: improving prediction of the minority class in boosting. Paper presented at: Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery; 2003:107-119; Springer, New York, NY.
15. Fan J, Fan Y. High dimensional classification using features annealed independence rules. *Ann Stat*. 2008;36(6):2605.
16. Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann Stat*. 2010;38(6):3567-3604.

17. Mai Q, Zou H. The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*. 2013;100(1): 229-234.
18. Yu W, Park T. AucPR: an AUC-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC Genom*. 2014;15(10):1-12.
19. Tomasko L, Helms RW, Snapinn SM. A discriminant analysis extension to mixed models. *Stat Med*. 1999;18(10):1249-1260.
20. Marshall G, Barón AE. Linear discriminant models for unbalanced longitudinal data. *Stat Med*. 2000;19(15):1969-1981.
21. De La Cruz-Mesia R, Quintana FA. A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hCG profiles. *Biostatistics*. 2007;8(2):228-238.
22. Arribas-Gil A, Cruz DIR, Lebarbier E, Meza C. Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. *Biometrics*. 2015;71(2):333-343.
23. Wilson D, Peters R, Ritchie K, Ritchie CW. Latest advances on interventions that may prevent, delay or ameliorate dementia. *Ther Adv Chronic Dis*. 2011;2(3):161-173.
24. The Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimers Dement*. 2021;17(3):321-387.
25. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc*. 2005;100(470):577-590.
26. Li K, O'Brien R, Lutz M, Luo S, Initiative ADN, others . A prognostic model of Alzheimer's disease relying on multiple longitudinal measures and time-to-event data. *Alzheimers Dement* 2018; 14(5): 644–651.
27. Li K, Luo S. Dynamic prediction of Alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Stat Med*. 2019;38(24):4804-4818.
28. Karhunen K. *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*. Vol 37. Helsinki, Finland: Soumalainen Tiedeakatemia; 1947.
29. Loève M. Functions aleatoires du second ordre. In: Levy P, ed. *Processus Stochastique et Mouvement Brownien*. Paris, French: Gauthier-Villars; 1965:366-420.
30. Staniswalis JG, Lee JJ. Nonparametric regression analysis of longitudinal data. *J Am Stat Assoc*. 1998;93(444):1403-1418.
31. Yao F, Müller HG, Clifford AJ, et al. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*. 2003;59(3):676-685.
32. Goldsmith J, Greven S, Crainiceanu C. Corrected confidence bands for functional data using principal components. *Biometrics*. 2013;69(1):41-51.
33. Di CZ, Crainiceanu CM, Caffo BS, Punjabi NM. Multilevel functional principal component analysis. *Ann Appl Stat*. 2009;3(1):458.
34. Yan L, Dodier RH, Mozer M, Wolniewicz RH. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. Paper presented at: Proceedings of the 20th International Conference on Machine Learning (ICML-03); 2003:848-855.
35. Hu J, Yang H, Lyu MR, King I, Man-Cho SA. Online nonlinear AUC maximization for imbalanced data sets. *IEEE Trans Neural NetwLearn Syst*. 2018;29(4):882-895. doi:10.1109/TNNLS.2016.2610465
36. Ma S, Huang J. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*. 2005;21(24):4356-4362.
37. Ma S, Huang J. Combining multiple markers for classification using ROC. *Biometrics*. 2007;63(3):751-757.
38. Wang Z, YcI C, Ying Z, Zhu L, Yang Y. A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics*. 2007;23(20):2788-2794.
39. Zhao X, Dai W, Li Y, Tian L. AUC-based biomarker ensemble with an application on gene scores predicting low bone mineral density. *Bioinformatics*. 2011;27(21):3050-3055.
40. Zhou X, Chen B, Xie Y, Tian F, Liu H, Liang X. Variable selection using the optimal ROC curve: an application to a traditional Chinese medicine study on osteoporosis disease. *Stat Med*. 2012;31(7):628-635.
41. Calders T, Jaroszewicz S. Efficient AUC optimization for classification. Paper presented at: Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery; 2007:42-53; Springer, New York, NY.
42. Kotlowski W, Dembczynski K, Huellermeier E. Bipartite ranking through minimization of univariate loss. ICML; 2011.
43. Zhao P, Hoi SC, Jin R, YANG T. Online AUC maximization; 2011.
44. Lyu S, Ying Y. A univariate bound of area under ROC; 2018. arXiv preprint arXiv:1804.05981.
45. Agarwal S. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. Paper presented at: Proceedings of the 26th Annual Conference on Learning Theory; Vol. 30, 2013:338-353. PMLR, Princeton, NJ, USA.
46. Menon AK, Williamson RC. Bayes-optimal scorers for bipartite ranking. Paper presented at: Proceedings of the 27th Conference on Learning Theory; Vol. 35, 2014:68-106. PMLR, Barcelona, Spain.
47. Gao W, Zhou ZH. On the consistency of AUC pairwise optimization. Paper presented at: Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015). 2015;939-945. AAAI Press, Buenos Aires, Argentina.
48. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J Royal Stat Soc Ser B (Stat Methodol)*. 2006;68(1):49-67.
49. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011;39(5):1.
50. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optim Theory Appl*. 2001;109(3):475-494.
51. Krishnapuram B, Carin L, Figueiredo MA, Hartemink AJ. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Trans Pattern Anal Mach Intell*. 2005;27(6):957-968.
52. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *J Royal Stat Soc Ser B (Stat Methodol)*. 2008;70(1):53-71.
53. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1.

54. Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement*. 2005;1(1):55-66.
55. Jack CR Jr, Bernstein MA, Fox NC, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magnet Reson Imaging Offic J Int Soc Magnet Reson Med*. 2008;27(4):685-691.
56. Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol*. 2010;6(2):67-77.
57. Zhang J, Liu M, An L, Gao Y, Shen D. Landmark-based Alzheimer's disease diagnosis using longitudinal structural MR images. *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*. New York, NY: Springer; 2016:35-45.
58. Gavidia-Bovadilla G, Kanaan-Izquierdo S, Mataró-Serrat M, Perera-Lluna A, Initiative ADN. Early prediction of Alzheimer's disease using null longitudinal model-based classifiers. *PLoS One*. 2017;12(1):e0168011.
59. Long X, Chen L, Jiang C, Zhang L, Initiative ADN. Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PLoS One*. 2017;12(3):e0173372.
60. Huang M, Yang W, Feng Q, Chen W. Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease. *Sci Rep*. 2017;7(1):1-13.
61. Mosconi L. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. *Eur J Nucl Med Mol Imaging*. 2005;32(4):486-510.
62. Li Y, Rinne JO, Mosconi L, et al. Regional analysis of FDG and PIB-PET images in normal aging, mild cognitive impairment, and Alzheimer's disease. *Eur J Nucl Med Mol Imaging*. 2008;35(12):2169-2181.
63. Langbaum JB, Chen K, Caselli RJ, et al. Hypometabolism in Alzheimer-affected brain regions in cognitively healthy Latino individuals carrying the apolipoprotein E ε4 allele. *Arch Neurol*. 2010;67(4):462-468.
64. Biagioni MC, Galvin JE. Using biomarkers to improve detection of Alzheimer's disease. *Neurodegener Dis Manag*. 2011;1(2):127-139.
65. Mattsson N, Zetterberg H, Hansson O, et al. CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA*. 2009;302(4):385-393.
66. Fjell AM, Walhovd KB, Fennema-Notestine C, et al. CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. *J Neurosci*. 2010;30(6):2088-2101.
67. Niemantsverdriet E, Valckx S, Bjerke M, Engelborghs S. Alzheimer's disease CSF biomarkers: clinical indications and rational use. *Acta Neurol Belg*. 2017;117(3):591-602.
68. Lee JC, Kim SJ, Hong S, Kim Y. Diagnosis of Alzheimer's disease utilizing amyloid and tau as fluid biomarkers. *Exp Mol Med*. 2019;51(5):1-10.
69. Reuter M, Schmansky NJ, Rosas HD, Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*. 2012;61(4):1402-1418.
70. Fischl B, Salat DH, Busa E, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*. 2002;33(3):341-355.
71. Fischl B. FreeSurfer. *NeuroImage*. 2012;62(2):774-781.
72. Li K, Chan W, Doody RS, Quinn J, Luo S. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *J Alzheimers Dis*. 2017;58(2):361-371.
73. Lin J, Li K, Luo S. Functional survival forests for multivariate longitudinal outcomes: dynamic prediction of Alzheimer's disease progression. *Stat Methods Med Res*. 2020;30(1):99-111.
74. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts; 2018.
75. Ghazarian AL, Haim T, Sauma S, Katiyar P. National Institute on Aging seed funding enables Alzheimer's disease startups to reach key value inflection points. *Alzheimers Dement*. 2022;18(2):348-359.
76. Li Z, Xie W, Liu T. Efficient feature selection and classification for microarray data. *PLoS One*. 2018;13(8):e0202167.
77. James GM, Hastie TJ, Sugar CA. Principal component models for sparse functional data. *Biometrika*. 2000;87(3):587-602.
78. James GM. Generalized linear models with functional predictors. *J Royal Stat Soc Ser B (Stat Methodol)*. 2002;64(3):411-432.
79. James GM, Sugar CA. Clustering for sparsely sampled functional data. *J Am Stat Assoc*. 2003;98(462):397-408.

How to cite this article: Li Y, Hsu W-W, Li Y. A classification for complex imbalanced data in disease screening and early diagnosis. *Statistics in Medicine*. 2022;41(19):3679-3695. doi: 10.1002/sim.9442