A classification for complex imbalanced data

by

Yiming Li

M.S., University of Florida, 2016

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Abstract

Imbalanced classification has drawn considerable attention in the statistics and machine learning literature. Typically, traditional classification methods, such as logistic regression and support vector machine (SVM), 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 areas including modern health research, face recognition, and object identification, it is expected that imbalanced classification may encounter an additional level of difficulty that is imposed by such a complex data structure.

In this dissertation, a nonparametric classification approach has been proposed for binary imbalanced data in longitudinal and high-dimensional settings. Technically, the proposed approach involves two stages. The functional principal component analysis (FPCA) is first applied for feature extraction under the sparse and irregular longitudinal data structure. The proposed univariate exponential loss function coupled with group LASSO penalty is then adopted into the classification procedure in high-dimensional settings. Along with the improvement in AUC and sensitivity under imbalanced classification, the proposed approach also provides a meaningful feature selection for interpretation while enjoying a remarkable computational efficiency. Finally, the proposed method is illustrated with the real data of Alzheimer's disease, Pima Indians diabetes and Phoneme, and its empirical performance in finite sample size is extensively evaluated by simulations.

Furthermore, the proposed method has been extended to multi-class scenario for which those aforementioned complications become more challenging. To accommodate the dense longitudinal/functional data, the use of natural cubic spline is adopted for feature extraction and dimension reduction, instead of using the FPCA. Functional biomarkers are efficiently characterized by spline coefficients which are treated as features for subsequent classification procedure. With these transformed features, a novel exponential loss function is then proposed to cast the multi-class classification task as a single optimization problem. Coupled with the group LASSO penalty, the proposed approach is also capable of performing variable selection for each class individually. Besides that, a simple weight-adjusted margin can be easily incorporated into the proposed loss function to address the issue of imbalance in multi-class data. The overall empirical performance of the proposed framework is evaluated by simulations in both high- and low-dimensional settings. Finally, the proposed multi-class classification framework is illustrated using real data of Alzheimer's disease, gene expression, and human walking.

A classification for complex imbalanced data

by

Yiming Li

M.S., University of Florida, 2016

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:	Approved by:
Co-Major Professor Dr. Wei-Wen Hsu	Co-Major Professor Dr. Weixing Song

Copyright

Abstract

Imbalanced classification has drawn considerable attention in the statistics and machine learning literature. Typically, traditional classification methods, such as logistic regression and support vector machine (SVM), 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 areas including modern health research, face recognition, and object identification, it is expected that imbalanced classification may encounter an additional level of difficulty that is imposed by such a complex data structure.

In this dissertation, a nonparametric classification approach has been proposed for binary imbalanced data in longitudinal and high-dimensional settings. Technically, the proposed approach involves two stages. The functional principal component analysis (FPCA) is first applied for feature extraction under the sparse and irregular longitudinal data structure. The proposed univariate exponential loss function coupled with group LASSO penalty is then adopted into the classification procedure in high-dimensional settings. Along with the improvement in AUC and sensitivity under imbalanced classification, the proposed approach also provides a meaningful feature selection for interpretation while enjoying a remarkable computational efficiency. Finally, the proposed method is illustrated with the real data of Alzheimer's disease, Pima Indians diabetes and Phoneme, and its empirical performance in finite sample size is extensively evaluated by simulations.

Furthermore, the proposed method has been extended to multi-class scenario for which those aforementioned complications become more challenging. To accommodate the dense longitudinal/functional data, the use of natural cubic spline is adopted for feature extraction and dimension reduction, instead of using the FPCA. Functional biomarkers are efficiently characterized by spline coefficients which are treated as features for subsequent classification procedure. With these transformed features, a novel exponential loss function is then proposed to cast the multi-class classification task as a single optimization problem. Coupled with the group LASSO penalty, the proposed approach is also capable of performing variable selection for each class individually. Besides that, a simple weight-adjusted margin can be easily incorporated into the proposed loss function to address the issue of imbalance in multi-class data. The overall empirical performance of the proposed framework is evaluated by simulations in both high- and low-dimensional settings. Finally, the proposed multi-class classification framework is illustrated using real data of Alzheimer's disease, gene expression, and human walking.

Table of Contents

List of Figures

List of Tables

Acknowledgments

It has been an arduous but rewarding journey since I began graduate school. I would like to extend thanks to many people who have assisted and influenced me on this journey.

First and foremost, I would like to express my sincere appreciation to my major advisor, Dr. Wei-Wen Hsu, for his invaluable guidance, encouragement and support. His knowledge and ethics, as well as his humility and wisdom, have had a profound impact on me. Without him, I could not complete my doctoral study.

Besides, I would like to express my gratitude to Dr. Weixing Song for being willing to serve as my co-major advisor, offering me insightful advice, and assisting me with all the paperwork. Moreover, I would also like to thank Dr. Gyuhyeong Goh and Dr. Shing I. Chang for serving on my committee and providing constructive feedback on my dissertation.

In addition, I would like to extend my thanks to Dr. Christopher I. Vahl, Dr. Haiyan Wang, Dr. Abby Jager and Dr. Juan Du for their assistance and excellent courses during my time at K-State. Thanks are also extended to Bonnie Messmer and Jo Blackburn for their assistance with the administrative paperwork. Additionally, I'd like to thank all my friends for their generosity and support throughout the years. Thank the Department of Statistics for all the financial support throughout my PhD studies, and thank the Coyne and Fryer families for providing me with the scholarships.

Last but not the least, I would like to express my deepest appreciation to my parents for their support and encouragement. I'd also like to thank my girlfriend, Qing, for her affection and support.

Chapter 1

Introduction

Imbalanced data are ubiquitous in real-world applications, such as medical diagnosis, object identification and image classification. Generally, traditional classifiers assume a balanced class distribution and target to minimize the overall misclassification rate. However, these methods usually perform poorly on imbalanced data and often misclassify instances from the minority class as ones from the majority class, thus 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. The classification of imbalanced data is even more challenging when the real data structure is complex, for example, in high-dimensional and longitudinal settings. In many disease screening and early diagnosis studies, longitudinal and/or high-dimensional data are common and 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. 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 (Japkowicz, 2000) or oversampling the minority class (Chawla et al., 2002; Batista et al., 2004). 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 (Sun et al., 2007, 2009). For instance, different penalties are assigned to different classes in the Support Vector Machine (SVM)-based classifiers (Lin et al., 2002). 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 (Margineantu, 2002; Zadrozny et al., 2003); 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 algorithm in these methods is basically centered around the combination of several simple classifiers/approaches in order to modify the training data sets for better prediction (Chawla et al., 2003; Seiffert et al., 2009; Wang and Japkowicz, 2010; Galar et al., 2013; Wan et al., 2014; Díez-Pastor et al., 2015).

To address the classification for high-dimensional data, several approaches have been proposed over the past decades. For example, Fan and Fan (2008) proposed the features annealed independence rules (FAIR) to select the most important features via a two-sample t-test. Fan and Song (2010) established a maximum-marginal-likelihood-type approach for feature screening. Mai and Zou (2013) 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 (2014) 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. (1999) and Marshall and Barón (2000) proposed a modified classical linear discriminant analysis using mixed-effects models to accommodate the over-time underlying associations. De La Cruz-Mesia and Quintana (2007) 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. (2015) considered a semiparametric linear mixed-effects model (SLMM) and proposed a unified estimation procedure based on a penalized EM-type algorithm. However, these methods 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. In this dissertation, we propose a two-stage approach to overcome these challenges in classification for complex imbalanced data. In the first stage, the techniques of functional principal component analysis (FPCA) are employed for feature extraction from longitudinal biomarkers. In other words, longitudinal data are analyzed via FPCA with a significant reduction in the longitudinal dimension, and then major principal components are treated as features for subsequent classification procedure. In the second stage, an univariate exponential loss is proposed to approximate the empirical area under the receiving operator characteristic (ROC) curve. The ROC analysis has been attracting attention in binary classification because it takes into account both true positive rate (TPR) and false positive rate (FPR) through a single summary measure, i.e., the area under ROC curve (AUC), and is invariant to the class distribution. Coupled with the group LASSO penalty, simultaneous model estimation and feature selection can be achieved efficiently by using the block-wise coordinate descent algorithm. The proposed method is illustrated on the binary data of Alzheimer's disease, Pima Indians diabetes, and Phoneme, and its empirical performance in finite sample sizes is extensively evaluated by simulations.

We further extend the proposed method to multi-class imbalanced data, where the classification is indeed more difficult as there could be multiple minority and majority classes. A particular class can be majority and minority concurrently, mainly depending on the sample size of the class it is compared to. Unfortunately, those methods stated previously for binary imbalanced data may not be generalized to multi-class scenario. For instance, for data-based approaches, the process of determining the sampling rate for each class can be tricky and the search space is much larger compared to the binary case (Fernández-Navarro et al., 2011). It is often complicated to select appropriate inductive biases for multiple majority and minority classes for algorithm-based approaches (Zhou and Liu, 2005).

The most commonly used approach to dealing with the multi-class imbalanced data is to employ the decomposition strategies (Galar et al., 2011; Kuncheva, 2014). This type of approach generally divides the original dataset into multiple binary subsets where the class imbalance will be accommodated separately. One popular method of this type of approach is the One-Versus-One (Knerr et al., 1990; Kreissel, 1999) method which constructs pairwise binary classifiers between classes. However, this method only uses partial data for the construction of each classifier and the computation is often intensive when the total number of classes is large. Another similar method is the One-Versus-All (Bottou et al., 1994) method in which a binary classifier is developed for each class. But this One-Versus-All approach is often unclear whether the output scores from different classifiers are on the comparable scales (Smola and Schölkopf, 1998). Besides these methods which simply reduce a multiclass classification problem to multiple binary tasks, several margin-based support vector machine (SVM) variants have been proposed, where the multi-class learning task is cast as a single optimization problem (Dogan et al., 2016). However, their algorithms are less efficient due to the non-convex structure in the loss function. Also, large samples are often required for model training, which may not be satisfied in typical clinical research. Some other studies have proposed the multi-class analogs of AUC which may suffer from either high computational complexity or ambiguous interpretation of the volume under the constructed ROC surface (Tang et al., 2011; Kleiman and Page, 2019).

To cope with these challenges associated with the classification for multi-class imbalanced data in a complex structure, we first apply the technique of natural cubic spline to dense longitudinal/functional data. Specifically, functional biomarkers are efficiently characterized by spline coefficients which can be treated as features for subsequent classification task. It is worth mentioning that the FPCA can also be used for feature extraction but its efficiency heavily depends on the longitudinal dimension of data. The extraction process via FPCA could be extremely computationally intensive if the dimensionality is ultra high. After the dimension reduction of longitudinal profiles in the first stage, we then consider to integrate a weight-adjusted margin into our novel exponential loss function for multi-class data in the second stage. As a result, the multi-class classification task becomes a single optimization problem. Lastly, we illustrate the proposed multi-class framework using simulations and real data analyses, such as early detection and conversion prediction on Alzheimer's disease data, tumor discrimination on gene expression data, and individual recognition on human gait data.

The rest of the dissertation is organized as follows. In Chapter 2, the FPCA is briefly introduced and our AUC-type classification framework for binary imbalanced data is proposed. For the real applications, we illustrate our method with the real data of Alzheimer's disease, Pima Indians diabetes and Phoneme. The empirical performance is evaluated by extensive simulations in finite sample size. In Chapter 3, we extend the proposed binary classifier to multi-class data by using a novel weight-adjusted exponential loss function. The feature selection can be conducted simultaneously for all classes through the partial Newton steps. Lastly, conclusions and discussions are provided in Chapter 4.

Chapter 2

The proposed AUC-type classification framework

2.1 Motivating example: Alzheimer's disease

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 (Wilson et al., 2011) 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 65yr+) is approximately 11% (Association, 2021), 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. Additionally, we also observe that some high-dimensional longitudinal biomarker data, such as brain imaging data, are collected irregularly and sparsely in the ADNI study, which further escalates the challenge of classification as we mentioned earlier.

The goals of this study include: (1) extract features from longitudinal data for subsequent classification procedure, (2) deal with the highly skewed class distribution to improve the classification performance, especially in terms of sensitivity and AUC, and (3) identify the most significant biomarkers that are associated with the progression of AD. To achieve these goals, the functional principal component analysis (FPCA) is employed in the first stage to accommodate the irregularly and sparsely longitudinal biomarkers for feature extraction. We then propose to use the univariate exponential loss to approximate the empirical AUC in the objective function to handle the class imbalance. Lastly, the variable selection is conducted by incorporating the group LASSO penalty into the proposed loss function.

2.2 Functional principal component analysis

To perform a functional principal component analysis (FPCA) on irregular and sparse longitudinal data, we adopt a version of FPCA proposed by Yao et al. (2005), 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 (Li et al., 2018; Li and Luo, 2019) and used for classification subsequently.

Assume that $M_{ij}(t)$ is the longitudinal trajectory of the j^{th} predictor of the i^{th} subject with $t \in \{1, ..., T_i\}$. 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 $\{\lambda_{jv}\}_{v=1,...,\infty}$ are nonincreasing eigenvalues, i.e., $\lambda_{j1} \geq ... \geq \lambda_{j\infty} \geq 0$, and $\{\phi_{jv}\}_{v=1,...,\infty}$ are the corresponding orthonormal eigenfunctions.

Using the Karhunen-Loève (KL) expansion (Karhunen, 1947; Loeve, 1948), $M_{ij}(t)$ can be expressed as

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

where $\{\xi_{ijv}\}_{v=1,\ldots,\infty}$ 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 $\sum_{v=1}^{\mathcal{V}} \lambda_{jv} / \sum_{v=1}^{\infty} \lambda_{jv}$ $\geq$ 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

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

where $\varepsilon_{ij}(t)$ is the measurement error with mean zero and variance σ^2. By applying PACE to the j^{th} 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 Staniswalis and Lee (1998) and Yao et al. (2003). 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 (Goldsmith et al., 2013). Finally, FPC scores $\{\xi_{ijv}\}^{'}$s for the i^{th} subject are estimated as follows:

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

where $\hat{\mu}_{ij} = \{\hat{\mu}_j(t)\}_{t=1,\ldots,T_i}$ and $\hat{\phi}_{ijv} = \{\hat{\phi}_{jv}(t)\}_{t=1,\ldots,T_i}$ 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 $t,\ t^{'} \in \{1, ..., T_i\}$.

Note that all these FPC scores can be obtained by using the `fpca.sc` function (Staniswalis and Lee, 1998; Di et al., 2009; Goldsmith et al., 2013) 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 paper. After obtaining theses FPC scores, a classification procedure can be applied subsequently.

2.3 Empirical AUC and its surrogate losses

The area under the receiver operating characteristic (ROC) curve, i.e., 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 (Yan et al., 2003; Hu et al., 2017).

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 r^{th} and s^{th} subjects in the health and disease groups, respectively, where $r = 1, ..., n_h$, $s = 1, ..., n_d$, 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 (Ma and Huang, 2005; Ma and Huang, 2007; Wang et al., 2007; Zhao et al., 2011b; Zhou et al., 2012). However, it usually necessitates pairwise comparisons between positive and negative instances, resulting in quadratic computational complexity (Calders and Jaroszewicz, 2007; Kotlowski et al., 2011; Zhao et al., 2011a; Lyu and Ying, 2018). 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 (Kotlowski et al., 2011; Agarwal, 2013; Menon and Williamson, 2014). Besides that, Gao and Zhou (2015) 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.

2.4 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:

$$\ell(\beta) = \sum_{i=1}^{N} e^{-y_i x_i^T \beta}, \tag{2.4.1}$$

where x_i is a vector containing all FPC scores of i^{th} subject, y_i is the corresponding response with binary outcomes, i.e., $y_i = 1$ if positive and $y_i = -1$ if negative (Menon and Williamson, 2014), 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. Due to high-dimensionality, we adopt the group lasso penalty proposed by Yuan and Lin (2006) to accommodate the grouping structure and perform group-feature selection. The objective function can be written as:

$$\ell_\tau(\beta) = \frac{1}{N} \sum_{i=1}^{N} e^{-y_i x_i^T \beta} + \tau \sum_{g=1}^{G} \sqrt{p_g} ||\beta_g||_2, \tag{2.4.2}$$

where β_g is a coefficient vector corresponding to the g^{th} grouped feature, p_g is the number of FPC scores within the g^{th} 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 $\mathcal{D}$-fold cross-validation with empirical AUC or univariate exponential loss, which are indeed equivalent in terms of classification performance. By the ease of interpretation of AUC, we use empirical AUC as criterion for all simulations and real data analyses throughout this study.

Regarding the choice of $\mathcal{D}$, it generally involves a trade-off between bias and variance. To be more precise, a large value of $\mathcal{D}$ typically results in small bias but large variance when evaluating the model performance, whereas a small value of $\mathcal{D}$ results in relatively large bias but small variance. The most commonly used values for $\mathcal{D}$ are $\mathcal{D} = 3$, 5, or 10. Considering the small sample size in the disease group under data imbalance, we adopt a 5-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 $\mathcal{D}$-fold cross-validation based on the sample size and the severity of imbalance.

To solve for the β that minimizes Equation (2.4.2), we employ a quadratic approximation which is similar to that in Simon et al. (2011). Let $m = X\beta$, where $X = [x_1, x_2, \ldots, x_N]^T$ is the design matrix, and $\dot{\ell}(\beta)$, $\ddot{\ell}(\beta)$, $\ell'(m)$, $\ell''(m)$ be the gradient and Hessian of the loss function in Equation (2.4.1) with respect to β and m, respectively. Using a second-order Taylor expansion centered at the initial value $\tilde{\beta}$, Equation (2.4.1) becomes:

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

where $\tilde{m} = X\tilde{\beta}$, $z(\tilde{m}) = \tilde{m} - \ell''(\tilde{m})^{-1}\ell'(\tilde{m})$, 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

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

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 $\{\beta_g\}_{g=1,\ldots,G}$ can be estimated iteratively by the block coordinate descent algorithm presented by Yuan and Lin (2006).

Specifically, to solve for the coefficients vector β_q for the q^{th} grouped feature, we first compute the corresponding first derivative of $L_\tau(\beta)$ as:

$$\frac{\partial L_\tau(\beta)}{\partial \beta_q} = -\frac{1}{N} X_q^T \ell''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq q} X_g \beta_g - X_q \beta_q \right) + \tau \sqrt{p_q} s_q, \tag{2.4.3}$$

where X_g and X_q are the data matrices corresponding to the g^{th} and q^{th} grouped features respectively, p_q is the group size of q^{th} grouped feature, and

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

Next, by setting Equation (2.4.3) to zero, we can obtain $\hat{\beta}_q$. Specifically, when $\beta_q = \mathbf{0}$, we can get:

$$\left\| \frac{1}{N} X_q^T \ell''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq q} X_g \beta_g \right) \right\|_2 \leqslant \tau \sqrt{p_q}, \tag{2.4.4}$$

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

$$\hat{\beta}_q = \left[\frac{1}{N} X_q^T \ell''(\tilde{m}) X_q + \frac{\tau \sqrt{p_q}}{||\beta_q||_2} \cdot I \right]^{-1} \cdot \left[\frac{1}{N} X_q^T \ell''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq q} X_g \beta_g \right) \right]. \tag{2.4.5}$$

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

Algorithm 1

```
Step 1. Initialize $\tilde{\beta}$, and compute $\tilde{m}$, $\ell'(\tilde{m})$, $\ell''(\tilde{m})$, and $z(\tilde{m})$.
Step 2. For $q = 1, ..., G$, if Equation (2.4.4) holds, $\hat{\beta}_q$ is set to $\mathbf{0}$; otherwise, $\hat{\beta}_q$ is updated
        using Equation (2.4.5).
Step 3. Set $\tilde{\beta} = \hat{\beta}$, and compute $\tilde{m}$, $\ell'(\tilde{m})$, $\ell''(\tilde{m})$, 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 (2001). 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 $\tilde{\beta} = (0.001, ..., 0.001)$ as we used in this study.

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 by Krishnapuram et al. (2005), Meier et al. (2008) and Friedman et al. (2010). 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.

2.5 Real applications

In this section, the classification performance of the proposed method is evaluated on three datasets, one with a longitudinal data structure in high-dimensional setting, i.e., Alzheimer's disease data, one with a cross-sectional data structure in low-dimensional setting, i.e., Pima Indians diabetes data, and the other one with a longitudinal data structure in low-dimensional setting, i.e., Phoneme data.

2.5.1 Alzheimer's Disease data

Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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 Mueller et al. (2005) and Jack Jr et al. (2008). In the ADNI dataset, participants are labeled with: cognitively normal (CN), MCI, or AD based on a series of assessments at the initial visits. These are also their states at baseline. It is expected to have repeated evaluations conducted subsequently at a six-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 2.1. 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 2.2.

Table 2.1: Demographic characteristics of selected subjects

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

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 (Frisoni et al., 2010; Zhang et al., 2016; Gavidia-Bovadilla et al., 2017; Huang et al., 2017; Long et al., 2017). Fluorodeoxyglucose positron emission tomography (FDG-PET) is able to provide the estimates of cerebral metabolic rates of glucose, thus revealing the pattern of

Table 2.2: 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

regional hypometabolism which is a prominent hallmark of AD (Mosconi, 2005; Li et al., 2008; Langbaum et al., 2010; Biagioni and Galvin, 2011). 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 (Mattsson et al., 2009; Fjell et al., 2010; Niemantsverdriet et al., 2017; Lee et al., 2019).

Brain imaging data preprocessing

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 which is an open-source software suite and freely available at *FreeSurferWiki* (https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki). The complete procedure used to preprocess the MRI data is summarized in Figure 2.1. 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 paper of Reuter et al. (2012). There are 319 biomarkers in total generated by Freesurfer v6.0, 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

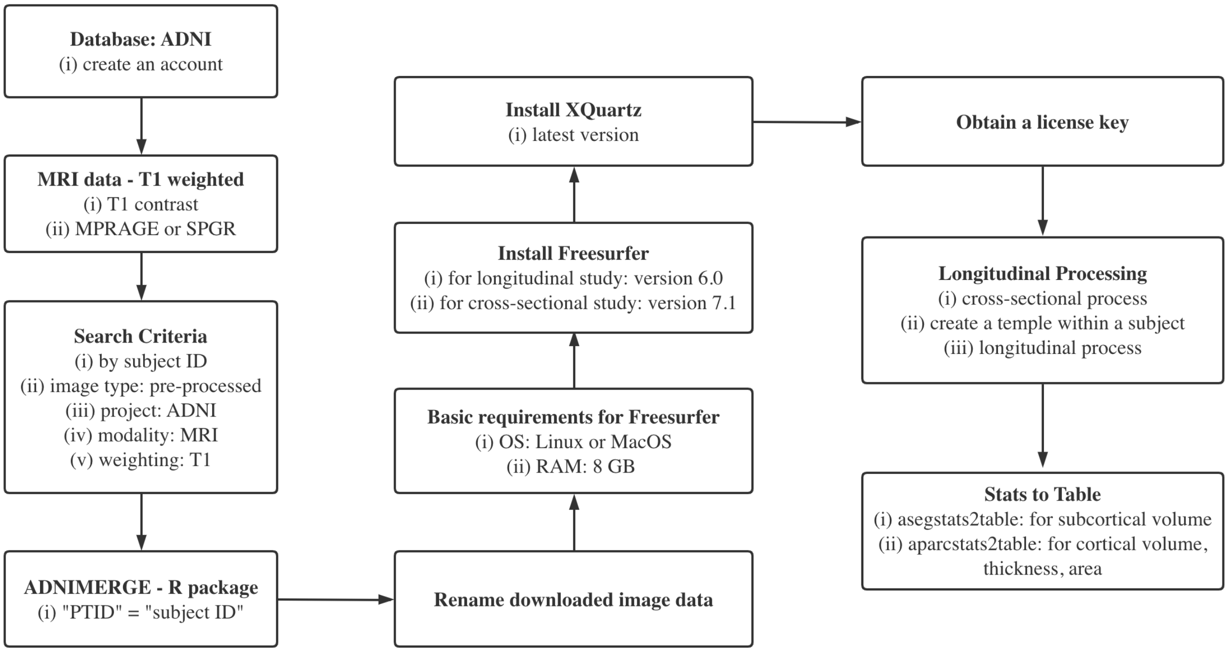
Database: ADNI
(i) create an account
MRI data - T1 weighted
(i) T1 contrast
(ii) MPRAGE or SPGR
Search Criteria
(i) by subject ID
(ii) image type: pre-processed
(iii) project: ADNI
(iv) modality: MRI
(v) weighting: T1
ADNIMERGE - R package
(i) "PTID" = "subject ID"
Rename downloaded image data
Basic requirements for Freesurfer
(i) OS: Linux or MacOS
(ii) RAM: 8 GB
Install Freesurfer
(i) for longitudinal study: version 6.0
(ii) for cross-sectional study: version 7.1
Install XQuartz
(i) latest version
Obtain a license key
Longitudinal Processing
(i) cross-sectional process
(ii) create a temple within a subject
(iii) longitudinal process
Stats to Table
(i) asegstats2table: for subcortical volume
(ii) aparcstats2table: for cortical volume, thickness, area

Figure 2.1: The procedure of preprocessing MRI data

range of subcortical structures (Fischl et al., 2002; Fischl, 2012).

Covariates and longitudinal biomarkers

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 (Li et al., 2017; Li and Luo, 2019; Lin et al., 2020): 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 2.2 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 is excluded. But for AD patients, we use the data before their first diagnosis of AD in order to train the model only based

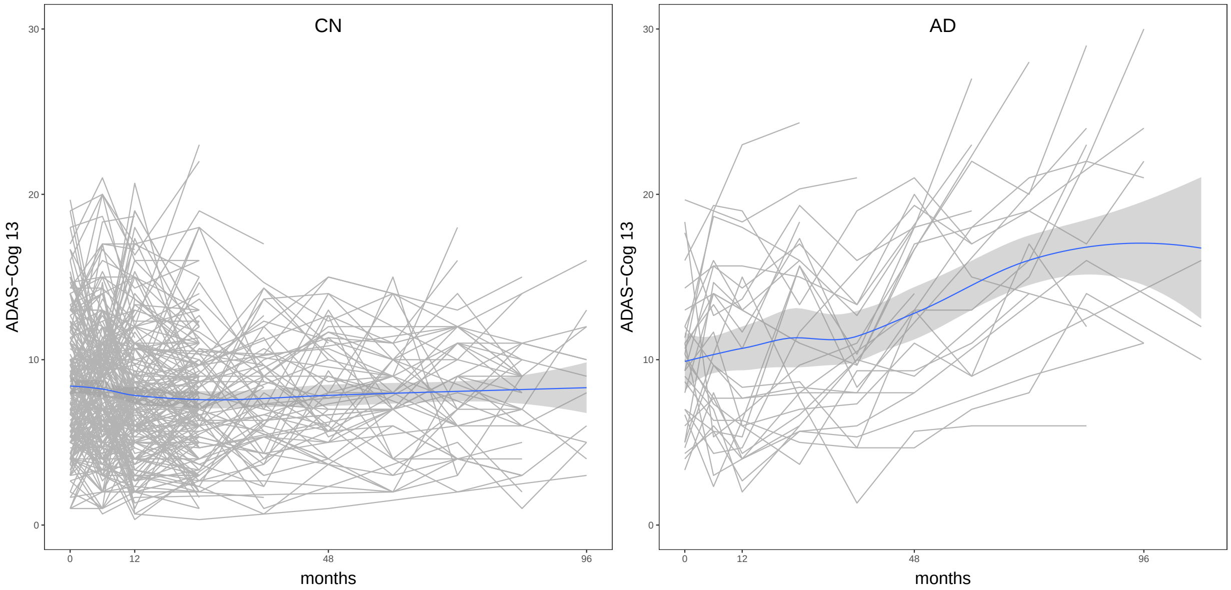

Figure 2.2: Longitudinal trajectories of ADAS-Cog 13 for cognitively normal subjects

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 2.3 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 (Hyndman and Athanasopoulos, 2018). To deal with these longitudinal biomarkers, the PACE algorithm proposed by Yao et al. (2005) 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 support vector machine (SVM) with linear kernel are also conducted with this ADNI dataset. The results based on 500 Monte Carlo replicates are given in Table 2.3. It is worth noting that the class distribution is highly imbalanced in this ADNI dataset (i.e., 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

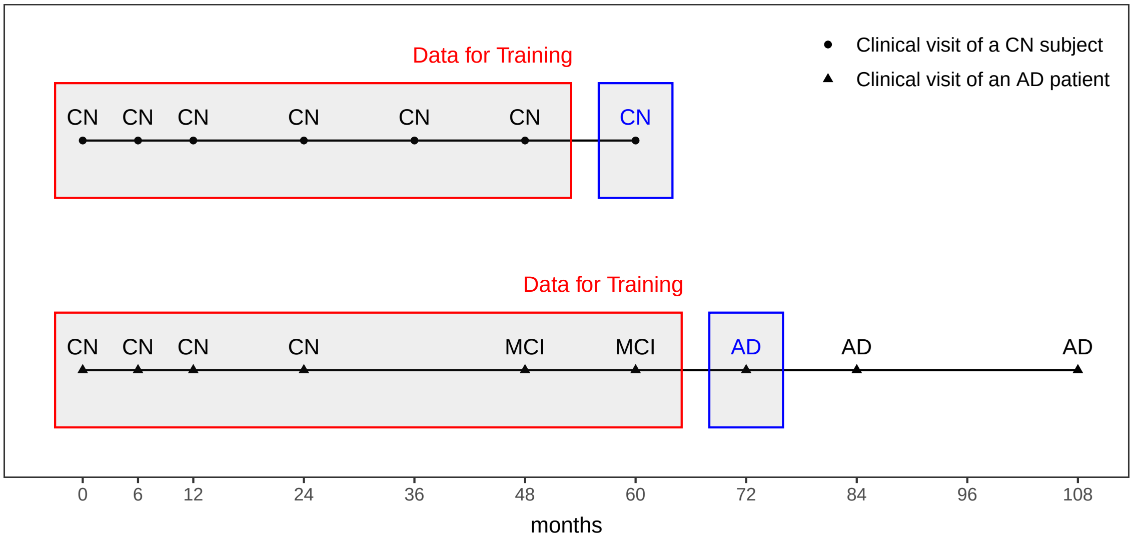

Figure 2.3: Clinical diagnosis of a CN subject or an AD patient over time. The red box represents the data used for model training. The blue box represents the final diagnosis used as the model outcome.

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 2.3, 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. Lastly, 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, right medial orbitofrontal cortex, right supramarginal gyrus, right pericalcarine cortex in cortical thicknesses. As a demonstration, the selected subcortical volumes which are considered to be correlated with the progression of AD are presented in Figure 2.4. Albeit interesting, more thorough investigations from the view of neuroscience are strongly encouraged before coming to any further conclusions.

Table 2.3: 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	.601(.297)	.999(.001)	.946(.066)
(n_h=166, n_d=21)	Specificity	.999(.001)	.999(.001)	.973(.035)
	Accuracy	.956(.033)	.999(.001)	.970(.035)
	AUC	.918(.167)	.999(.001)	.976(.033)
Test Set	Sensitivity	.362(.199)	.441(.154)	.790(.145)
(n_h=71, n_d=9)	Specificity	.996(.008)	.980(.015)	.880(.094)
	Accuracy	.925(.022)	.919(.023)	.870(.084)
	AUC	.832(.147)	.854(.068)	.880(.091)

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

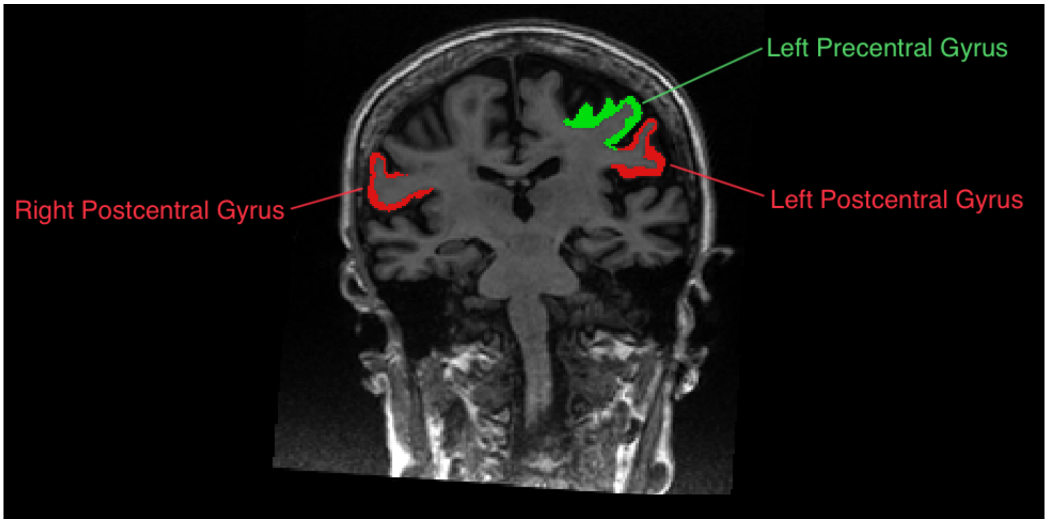

Figure 2.4: Selected subcortical volumes in coronal view.

2.5.2 Pima Indians Diabetes data

The data used in this analysis are from a study led by the National Institute of Diabetes and Digestive and Kidney Diseases and publicly accessible from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). A subset of 768 subjects are chosen from the Pima Indian population near Phoenix, Arizona, of which 268 were diagnosed with diabetes, thus leading to an imbalanced class distribution. More details about this study and the eligibility criteria of subjects can be found in the paper of Smith et al. (1988). It is noted that all these subjects are females over 21 years old. To forecast the onset of diabetes, eight variables which are found to be common risk factors for diabetes are used in this analysis, including the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, serum insulin, body mass index (BMI), diabetes pedigree function and age (Ma and Huang, 2007). The demographic information of subjects in this dataset is summarized in Table 2.4.

Table 2.4: Demographic information of subjects in the diabetes dataset

		Age (years)		
Group	n	Range	Mean	Std. Dev
non-Diabetes	500	[21,81]	31.2	11.7
Diabetes	268	[21,70]	37.1	11.0

In our analysis, 70% of subjects are randomly selected as a training set and the remaining 30% are used as a test set. All of the continuous predictors are normalized with zero mean and unit variance. The optimal tuning parameter in our model is determined by five-fold cross-validation. To evaluate the classification performance, four widely used measures are utilized, i.e., accuracy, sensitivity, specificity and AUC. Besides that, we include another three metrics which are often employed in imbalanced classification, i.e., precision, G-mean and F-measure (Yeh et al., 2016; Liu et al., 2020; Tanha et al., 2020). For the purpose of comparison, the results of the proposed method, logistic regression and linear SVM based on 500 Monte Carlo replicates are presented in Table 2.5.

Table 2.5: Classification results (S.E.) for Pima Indians Diabetes data with logistic regression, linear SVM and the proposed method based on 500 Monte Carlo replicates

		Logistic	linear SVM	Proposed
Training Set	Sensitivity	.581(.021)	.574(.024)	.750(.046)
(n_h=350, n_d=188)	Specificity	.886(.007)	.889(.010)	.758(.047)
	Accuracy	.779(.010)	.779(.010)	.755(.018)
	AUC	.841(.010)	.840(.010)	.812(.012)
Test Set	Sensitivity	.569(.050)	.562(.048)	.725(.072)
(n_h=150, n_d=80)	Specificity	.880(.028)	.882(.026)	.741(.057)
	Accuracy	.772(.022)	.771(.022)	.735(.027)
	AUC	.830(.023)	.829(.022)	.802(.027)

Logistic: logistic regression; linear SVM: support vector machine with linear kernel

(n_h, n_d): number of subjects in the non-diabetes and diabetes groups respectively

The accuracy and AUC of the three methods are close and comparable in both training and test sets. However, the proposed method achieves a much higher sensitivity than logistic regression and SVM while retaining an adequate specificity. Moreover, the proposed approach performs slightly better than the other two techniques in terms of F-measure and G-mean, which indicates its strength in dealing with the class imbalance.

It should be noted that the proposed classification in Section 2.4 has been modified to model this dataset. Due to the nature of cross-sectional data, there is no need to consider the grouping structure among features. Consequently, we set all group sizes $\{p_g\}_{g=1,..,G}$ in the penalty term to one, where G denotes the total number of groups, i.e. the number of features for this dataset.

2.5.3 Phoneme data

In this section, we illustrate our method with the phoneme data which are formed by selecting phonemes based on digitized speech from the Texas Instruments/Massachusetts Institute of Technology (TIMIT) database (available at https://web.stanford.edu/∼hastie/ElemStatLearn). This dataset consists of 4509 speech frames of 32 ms duration from continuous speech and each frame represents one of five phonemes transcribed as follows: 'sh' as in 'she', 'dcl' as in 'dark', 'iy' as the vowel in 'she', 'aa' as the vowel in 'dark', and 'ao' as the first vowel in 'water' (Hastie et al., 1995; Ferré and Villa, 2006; Shin, 2008; Delaigle and Hall, 2012). In other words, the data used in our analysis only have one time-series predictor with 4509 observations for five classes. A log-periodogram which is frequently used to cast speech data in a form suitable for speech recognition is then computed for each speech frame, thus forming a functional profile of length 256.

In this analysis, a binary classification problem is considered. The proposed longitudinal framework is applied to identifying whether a given speech frame is 'aa' or not. Figure 2.5 shows the curves of log-periodograms for 10 randomly selected speech frames from class 'aa' and 'non-aa' (i.e. 'ao', 'dcl', 'iy' and 'sh'), respectively. There are 4509 log-periodograms in this dataset, of which 695 are labeled as 'aa', thus leading to a classification with imbalanced class distribution. Model is fitted on the training data which are randomly selected and contain 70% of all instances, and then performance is assessed on the test data which contain the remaining 30% of instances. The optimal tuning parameter is selected by five-fold cross-validation using AUC as the criterion. Table 2.6 provides the classification results for our method, logistic regression with L_1 penalty and SVM with linear kernel based on 200 Monte Carlo replicates. The results of AUC and accuracy for three methods are very close and comparable. However, our proposed framework outperforms L_1 logistic regression and linear SVM in terms of sensitivity with slight compromise in specificity, which indicates the superiority of our method in imbalanced classification.

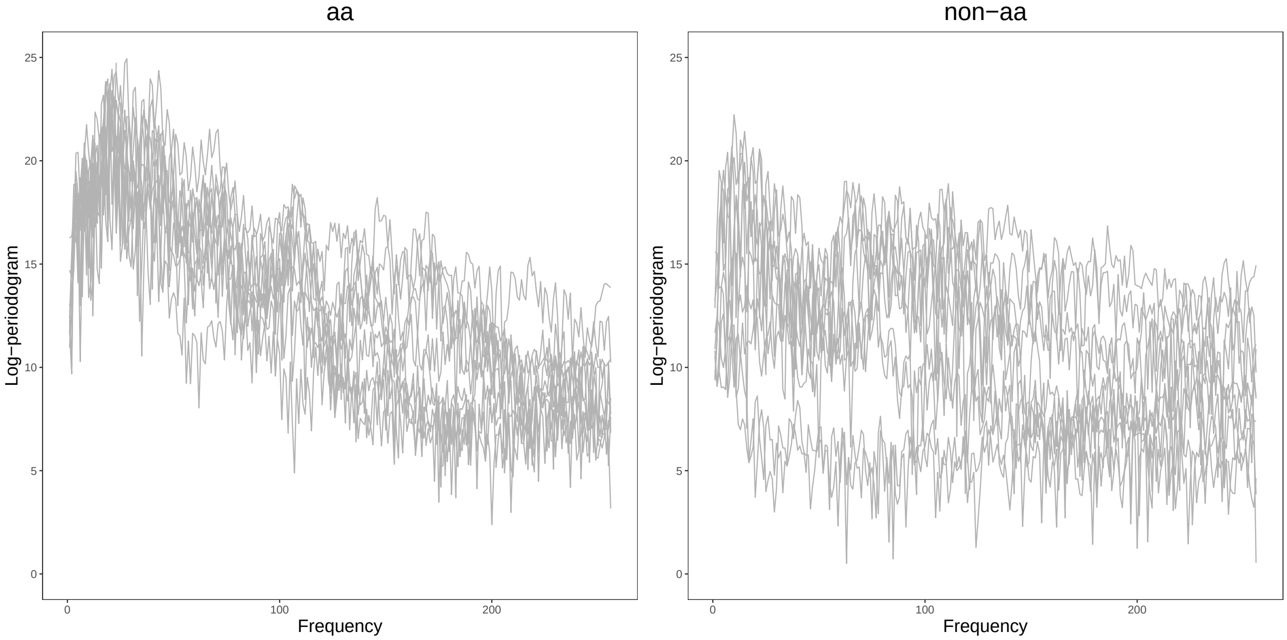

Figure 2.5: Curves of log-periodograms for 10 randomly selected speech frames from class ‘aa’ and ‘non-aa’

Table 2.6: Classification results for Phoneme data with L_1 Logistic, linear SVM and the Proposed Method based on 200 Monte Carlo replicates

	L_1 Logistic	linear SVM	Proposed
Training Set			
$(n_{ctrl} = 2670, n_{case} = 487)$			
Sensitivity	.595(.016)	.608(.016)	.907(.016)
Specificity	.954(.002)	.953(.003)	.843(.016)
Accuracy	.898(.003)	.899(.003)	.853(.012)
AUC	.940(.002)	.939(.002)	.940(.002)
Test Set			
$(n_{ctrl} = 1144, n_{case} = 208)$			
Sensitivity	.592(.032)	.605(.031)	.896(.025)
Specificity	.954(.006)	.952(.007)	.842(.019)
Accuracy	.898(.007)	.899(.007)	.851(.014)
AUC	.940(.005)	.939(.006)	.939(.005)

L_1 Logistic: logistic regression with L_1 penalty; linear SVM: SVM with linear kernel

n_{ctrl}: number of observations for control; n_{case}: number of observations for case

2.6 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 (i.e., $t \in \{0, 1, 2, 3, 4, 5, 6\}$ and $t = 0$ represents the baseline). We also perform other two popular methods (i.e., logistic regression and support vector machine) at various levels of class imbalance for comparison purposes.

2.6.1 Class memberships by a logistic regression model

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

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

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:

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

where β_j is the coefficient vector corresponding to the j^{th} longitudinal predictor and β_0 is the intercept which can be adjusted 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 2.7. In this setting, the performances of three methods are comparable in terms of AUC and accuracy. However, regardless of training or testing, remarkable lower sensitivities are observed in the methods of logistic regression and support vector machine 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 $\{\gamma_{\rho j}\}_{\rho=0,1,2}$ that correspond to the j^{th} predictor

Table 2.7: 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

	Imbalance ratio	$n_h/n_d = 3.2$			$n_h/n_d = 4.9$			$n_h/n_d = 6.1$		
		Logistic	SVM	Proposed	Logistic	SVM	Proposed	Logistic	SVM	Proposed
Training	Sensitivity	.689 (.061)	.681 (.071)	.873 (.035)	.618 (.081)	.594 (.100)	.884 (.039)	.548 (.108)	.483 (.156)	.896 (.041)
	Specificity	.941 (.013)	.945 (.015)	.856 (.034)	.965 (.009)	.970 (.012)	.867 (.034)	.981 (.007)	.987 (.008)	.882 (.038)
	Accuracy	.880 (.019)	.882 (.019)	.860 (.025)	.910 (.016)	.910 (.016)	.870 (.029)	.938 (.013)	.937 (.013)	.883 (.034)
	AUC	.933 (.016)	.932 (.016)	.932 (.016)	.940 (.017)	.937 (.018)	.937 (.017)	.948 (.017)	.945 (.019)	.945 (.018)
Test	Sensitivity	.672 (.065)	.658 (.071)	.833 (.059)	.594 (.087)	.564 (.097)	.828 (.074)	.507 (.111)	.435 (.147)	.819 (.087)
	Specificity	.933 (.021)	.935 (.021)	.840 (.039)	.959 (.015)	.964 (.016)	.857 (.037)	.975 (.012)	.981 (.013)	.871 (.040)
	Accuracy	.870 (.018)	.869 (.018)	.838 (.026)	.900 (.016)	.899 (.016)	.852 (.028)	.927 (.015)	.925 (.015)	.866 (.032)
	AUC	.923 (.017)	.922 (.018)	.921 (.018)	.929 (.019)	.927 (.019)	.927 (.020)	.934 (.020)	.932 (.020)	.931 (.021)

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

are generated randomly from truncated normal distributions (TN):

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

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 $\beta_j = (0, 0)$, $j \in \{6, ..., 500\}$. 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 2.8. In this setting, the performance of the proposed method is better than that of the other two approaches in terms of AUC

Table 2.8: 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

	Imbalance ratio	$n_h/n_d = 3.2$			$n_h/n_d = 4.9$			$n_h/n_d = 6.1$		
		Logistic	SVM	Proposed	Logistic	SVM	Proposed	Logistic	SVM	Proposed
Training	Sensitivity	.705 (.221)	.999 (.001)	.900 (.035)	.659 (.293)	.999 (.001)	.905 (.052)	.549 (.360)	.999 (.001)	.899 (.062)
	Specificity	.997 (.005)	.999 (.001)	.894 (.037)	.999 (.002)	.999 (.001)	.891 (.050)	.999 (.001)	.999 (.001)	.888 (.061)
	Accuracy	.928 (.053)	.999 (.001)	.896 (.032)	.942 (.049)	.999 (.001)	.893 (.047)	.938 (.049)	.999 (.001)	.889 (.058)
	AUC	.982 (.022)	.999 (.001)	.957 (.020)	.982 (.051)	.999 (.001)	.952 (.034)	.944 (.135)	.999 (.001)	.946 (.044)
Test	Sensitivity	.412 (.112)	.221 (.058)	.791 (.078)	.262 (.122)	.109 (.056)	.724 (.122)	.174 (.125)	.063 (.043)	.686 (.137)
	Specificity	.968 (.021)	.901 (.026)	.856 (.039)	.982 (.016)	.958 (.017)	.860 (.048)	.989 (.013)	.977 (.013)	.860 (.057)
	Accuracy	.836 (.025)	.740 (.023)	.841 (.031)	.859 (.021)	.813 (.020)	.837 (.037)	.874 (.019)	.848 (.018)	.835 (.044)
	AUC	.892 (.029)	.645 (.038)	.913 (.028)	.876 (.050)	.640 (.044)	.889 (.043)	.842 (.108)	.635 (.050)	.877 (.047)

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

and sensitivity. It seems that logistic regression and support vector machine tend to classify subjects into the majority class (i.e., the health group), thus resulting in low sensitivity. However, the proposed method achieves a better sensitivity with a little sacrifice of specificity and accuracy.

2.6.2 Class memberships by pre-determined 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:

$$\begin{aligned} U_{rj}^{H}(t) &= \mu_j^{H}(t) + b_{rj} + \varepsilon_{rj}(t), \ t \in \{0, 1, ..., 6\} \\ U_{sj}^{D}(t) &= \mu_j^{D}(t) + b_{sj} + \varepsilon_{sj}(t), \ t \in \{0, 1, ..., 6\} \end{aligned}$$

where $\mu_j^H(t)$ and $\mu_j^D(t)$ are the mean functions of the j^{th} longitudinal predictor for the r^{th} health and the s^{th} disease subject, respectively, b_{rj} and b_{sj} are the subject-specific random effects generated from $N(0, 1.5)$. The random errors $\varepsilon_{sj}(t)$ 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, i.e., $Y = 0$ 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 $t = 0, 1, ..., 6$, to reflect the progression of the disease. Three sets of $\{\gamma_{\rho j}\}_{\rho=0,1,2}$ 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, i.e., $\mu_j^H = \mu_j^D = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)^T,\ j \in \{26, ..., 500\}$. 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, i.e., $\mu_j^H = (c_j^H, c_j^H, ..., c_j^H)_{1\times 7}^T$, $j \in \{1, ..., 25\}$, 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 $\{\gamma_{\rho j}\}_{\rho=0,1,2}$ that correspond to the j^{th} predictor are randomly selected, for each Monte Carlo sample, from truncated normal distributions:

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

The simulation results are given in Table 2.9 and 2.10. Even under this data-generating mechanism, the proposed approach outperforms logistic regression and support vector machine 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 Table 2.9 and 2.10). 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 of logistic regression and support vector machine in the high-dimensional setting, also coming along with smaller standard errors. This result indeed indicates the stability of our approach in high-dimensional settings.

Table 2.9: Classification results (S.E.) with various sample sizes of health and disease groups in low-dimensional setting based on 500 Monte Carlo replicates

	Sample size	$(n_h, n_d) = (225, 75)$			$(n_h, n_d) = (257, 43)$			$(n_h, n_d) = (270, 30)$		
		Logistic	SVM	Proposed	Logistic	SVM	Proposed	Logistic	SVM	Proposed
Training	Sensitivity	.774 (.097)	.767 (.108)	.895 (.049)	.596 (.205)	.555 (.250)	.873 (.071)	.454 (.226)	.379 (.278)	.852 (.086)
	Specificity	.954 (.014)	.957 (.015)	.894 (.045)	.976 (.008)	.982 (.010)	.864 (.074)	.985 (.007)	.991 (.008)	.846 (.080)
	Accuracy	.909 (.032)	.909 (.032)	.895 (.042)	.922 (.029)	.921 (.032)	.866 (.070)	.932 (.021)	.930 (.023)	.846 (.079)
	AUC	.953 (.033)	.952 (.035)	.951 (.034)	.927 (.064)	.919 (.081)	.925 (.064)	.906 (.076)	.885 (.110)	.902 (.077)
Test	Sensitivity	.751 (.099)	.743 (.108)	.865 (.059)	.555 (.198)	.511 (.235)	.818 (.095)	.416 (.214)	.335 (.254)	.780 (.119)
	Specificity	.949 (.017)	.951 (.016)	.882 (.048)	.970 (.012)	.976 (.014)	.853 (.075)	.981 (.010)	.988 (.011)	.837 (.080)
	Accuracy	.899 (.033)	.898 (.032)	.878 (.044)	.911 (.029)	.909 (.030)	.848 (.071)	.925 (.019)	.922 (.020)	.832 (.077)
	AUC	.945 (.037)	.944 (.038)	.945 (.037)	.912 (.066)	.907 (.079)	.912 (.068)	.889 (.079)	.872 (.109)	.889 (.080)

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

Table 2.10: Classification results (S.E.) with various sample sizes of health and disease groups in high-dimensional setting based on 500 Monte Carlo replicates

	Sample size	$(n_h, n_d) = (225, 75)$			$(n_h, n_d) = (257, 43)$			$(n_h, n_d) = (270, 30)$		
		Logistic	SVM	Proposed	Logistic	SVM	Proposed	Logistic	SVM	Proposed
Training	Sensitivity	.825 (.121)	.999 (.001)	.927 (.033)	.782 (.234)	.999 (.001)	.921 (.039)	.395 (.369)	.999 (.001)	.911 (.053)
	Specificity	.992 (.007)	.999 (.001)	.924 (.032)	.999 (.001)	.999 (.001)	.918 (.035)	.999 (.001)	.999 (.001)	.904 (.053)
	Accuracy	.950 (.034)	.999 (.001)	.924 (.028)	.968 (.034)	.999 (.001)	.918 (.032)	.939 (.036)	.999 (.001)	.905 (.051)
	AUC	.987 (.013)	.999 (.001)	.972 (.015)	.993 (.024)	.999 (.001)	.968 (.019)	.842 (.224)	.999 (.001)	.956 (.034)
Test	Sensitivity	.657 (.063)	.457 (.056)	.861 (.053)	.389 (.113)	.192 (.056)	.793 (.076)	.131 (.121)	.064 (.047)	.709 (.113)
	Specificity	.967 (.015)	.930 (.016)	.892 (.031)	.987 (.009)	.981 (.008)	.895 (.037)	.997 (.004)	.994 (.005)	.885 (.054)
	Accuracy	.890 (.018)	.812 (.020)	.884 (.024)	.901 (.015)	.868 (.011)	.881 (.031)	.910 (.010)	.901 (.006)	.868 (.046)
	AUC	.947 (.016)	.832 (.029)	.951 (.017)	.922 (.032)	.807 (.037)	.931 (.026)	.777 (.183)	.762 (.049)	.897 (.042)

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

2.7 Discussion

In this work, we have developed a novel classification framework for imbalanced data under 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 (Ghazarian et al., 2021), 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 only 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, the imaging data may not always be available for each individual. 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 low-dimensional settings. Apart from the longitudinal data, the proposed method without FPC score extraction can be easily applied to cross-sectional imbalanced data.

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 (James et al., 2000; James, 2002; James and Sugar, 2003). 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. 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 (i.e., FPC scores) by PACE can still be performed even when missing values occur in the longitudinal profiles of subjects.

Finally, it is possible to extend our approach to incorporating other alternative surrogate loss functions for the approximation of the empirical AUC, such as square loss and squared hinge loss. Such an extension may potentially improve the classification performance and reduce the computational burden. Besides that, 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 work and require further investigations.

Chapter 3

The proposed multi-class classification framework

3.1 Motivating example: Human walking patterns

The extension of the proposed method in Chapter 2 to multi-class data is motivated by the application of walking patterns for person recognition. It is generally assumed that the human gait is a unique characteristic, thus providing critical and meaningful biometric information for identification (Kim and Kim, 2017; Wan et al., 2018; Nambiar et al., 2019). Typically, in the analysis of human walking patterns, the spatial-temporal gait data are often recorded longitudinally with multiple cycles collected for individuals. Therefore, gait-based person recognition can be viewed as a multi-class classification task. In other words, each individual is considered as one class and the gait cycles collected from each one are repeated observations. However, the number of gait cycles collected may vary across individuals, which can cause the issue of imbalance in data. Additionally, gait data may be acquired in high-dimensional settings, which further increases the difficulty of model estimation. It is worth mentioning that the identification of rehabilitation via human walking data has also been drawing attention in the literature (Baker, 2006; Horst et al., 2017). In this study, we mainly focus on the task of person recognition and have not conducted rehabilitation research due to the lack of sufficient gait data of individuals being in different conditions, for example, wearing ankle brace, knee brace or both.

Our goals in this study include: (1) transform the dense longitudinal/functional profiles into summary measures, (2) develop an efficient and unified classification framework for multi-class data, and (3) handle the issue of imbalance among classes. To achieve these goals, we propose a two-stage approach for multi-class imbalanced data. In the first stage, the techniques of natural cubic spline have been applied for feature extraction. As a result, a significant reduction in the functional dimension is conducted. In the second stage, we develop a novel exponential loss function which leads to an efficient optimization for the model estimation. With the incorporation of group LASSO penalty, variable selection can be performed simultaneously for all classes on the grouped level.

3.2 Natural cubic spline

To extract features from longitudinal/functional data, the natural cubic spline basis has been widely used, thus leading to a considerable dimension reduction (James et al., 2000; James, 2002; James and Sugar, 2003). Suppose that $f(Z)$ is a function on $[a, b]$ with a sequence of knots $\{\xi_i\}_{i=1,\dots,\lambda}$ which are defined as $a < \xi_1 < \xi_2 < \dots < \xi_\lambda < b$. It generally requires three conditions for such f to be a natural cubic spline: (1) f is a cubic polynomial on each of the intervals $\{(a, \xi_1), (\xi_1, \xi_2), \dots, (\xi_{\lambda-1}, \xi_\lambda), (\xi_\lambda, b)\}$, (2) f is continuous up to the second derivative, and (3) the second and third derivatives of f at both a and b are equal to zero. Typically, the third condition which is also called the natural boundary constraints guarantees that f is linear beyond the boundary knots a and b (Green and Silverman, 1993). A natural cubic spline with $\mathcal{K}$ knots can be represented by $\mathcal{K}$ basis functions as follows (Hastie et al., 2009):

$$h_1(Z) = 1,\ h_2(Z) = Z,\ h_{k+2}(Z) = \eta_k(Z) - \eta_{\mathcal{K}-1}(Z),\ k = 1, \dots, \mathcal{K} - 2,$$

where

$$\eta_k(Z) = \frac{(Z - \xi_k)_+^3 - (Z - \xi_{\mathcal{K}})_+^3}{\xi_{\mathcal{K}} - \xi_k},$$

and

$$(Z-\xi_k)_+ = \begin{cases} Z-\xi_k, & \text{if } Z-\xi_k > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Compared to other cubic splines such as the cubic B-spline, estimates generated from a natural cubic spline are more stable at the boundaries due to the boundary constraints (James et al., 2013).

Assume that $U_{ij}(t)$ is the observed j^{th} longitudinal biomarker of the i^{th} subject with $t \in \{1, ..., T_i\}$. Let $\{S_{jv}(t)\}_{v=1,...,\mathcal{V}}$ be a set of known basis functions of $U_{ij}(t)$. Then $U_{ij}(t)$ can be expressed as:

$$U_{ij}(t) = \sum_{v=1}^{\mathcal{V}} S_{jv}(t)\alpha_{jv},$$

where $\{\alpha_{jv}\}_{v=1,...,\mathcal{V}}$ is a set of spline coefficients corresponding to the j^{th} biomarker, and $\mathcal{V}$ is the dimension of expansion. The basis functions and $\mathcal{V}$ should be pre-specified before applying the spline representation. For the choice of basis functions, several widely used bases have been discussed (Ramsay and Silverman, 2006). For example, a Fourier basis is often employed for data with a periodic structure. A spline basis such as a cubic spline seems to be more appropriate for non-periodic data. In a natural cubic spline, $\mathcal{V}$ is usually determined by the number of inner knots. Typically, the curve is more smooth and flexible with more knots placed. A typical approach for determining the optimal value for $\mathcal{V}$ is to try out different numbers of knots. Another commonly adopted approach is to select the optimal $\mathcal{V}$ via cross-validation. In practice, a small value is often used for $\mathcal{V}$, such as $\mathcal{V} = 2$ or 3, to reduce the computational burden. Note that all the spline coefficients in this paper are obtained by using the `ns` function from the R package `splines`, and $\mathcal{V}$ is determined by setting a specific value for the degrees of freedom (df). With these extracted spline coefficients, a classification procedure can then be applied.

3.3 The proposed multi-class classification

Consider a training set $\{x_i, y_i\}_{i=1,...,N}$, where x_i is a p-dimensional vector consisting of all of the spline coefficients extracted from longitudinal biomarkers via natural cubic spline, y_i is the corresponding multi-class label, i.e., $y_i \in \{1, 2, ..., C\}$ with C being the total number of classes, and N denotes the total number of samples in this dataset. Motivated by the decision function constructed for multi-class support vector machines (SVM) (Weston and Watkins, 1998), we propose a novel exponential loss function for the classification of multi-class data as follows:

$$\ell(w) = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{c\neq y_i} e^{\left[-\left(x_i^T w_{y_i} - x_i^T w_c\right)\right]}\right\},\ c = 1, ..., C, \tag{3.3.1}$$

where w_{y_i} and w_c denote the vector of coefficients corresponding to the correct class of i^{th} sample and c^{th} class, respectively, and $w = [w_1, w_2, ..., w_C]^T$ is a $C \times p$ matrix.

To address the issue of class imbalance under the multi-class scenario, we further propose a simple class-based weight which can be easily integrated into Equation (3.3.1). Suppose that the sample size of c^{th} class is n_c, it is easy to get $N = \sum_{c=1}^{C} n_c$. Then the class-based weight ψ_c corresponding to c^{th} class can be computed as:

$$\psi_c = \frac{N}{Cn_c}.$$

Therefore, the weight-adjusted exponential loss is defined as:

$$\ell_\Delta(w) = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{c\neq y_i} e^{\left[-\left(x_i^T w_{y_i} - x_i^T w_c - \Delta_c\right)\right]}\right\}, \tag{3.3.2}$$

where $\Delta_c = \ln(\psi_c)$.

Assume that both w_{y_i} and w_c in Equation 3.3.2 are normalized (Liu et al., 2017; Wang et al., 2018; Deng et al., 2019), i.e., $||w_{y_i}|| = ||w_c|| = 1$, the proposed loss function can be rewritten as:

$$\ell(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{c\neq y_i} e^{\left[-||x_i||\left(\cos\theta_{y_i} - \cos\theta_c\right) + \Delta_c\right]}\right\},\ c = 1, ..., C,$$

where θ_{y_i} denotes the angle between x_i and w_{y_i}, and θ_c denotes the angle between x_i and w_c. It is easy to observe that minimizing $\ell(\theta)$ amounts to maximizing the angles between samples from different classes, thus encouraging intra-class compactness and enlarging inter-class separation. Hence, the minimization of Equation 3.3.2 is expected to separate distinct classes from each other due to the equivalence of $\ell(\theta)$ and $\ell_\Delta(w)$ and optimize the classification performance for multi-class data.

To avoid the computational challenges associated with the direct optimization of Equation (3.3.2), a partial quadratic approximation has been conducted by applying the partial Newton steps (Friedman et al., 2010). Thus, the coefficient vectors $\{w_c\}_{c=1,...,C}$ can be estimated iteratively for a single class at a time by cycling through C classes. Specifically, for a particular w_k, let $m = Xw_k$, where $X = [x_1, x_2, ..., x_N]^T$ is the design matrix, and $\dot{\ell}(w_k)$, $\ddot{\ell}(w_k)$, $\ell'(m)$, $\ell''(m)$ be the gradient and Hessian of the loss function in Equation (3.3.2) with respect to w_k and m, respectively. Using a second-order Taylor expansion centered at the initial value $\tilde{w}_k$, it is not difficult to show that:

$$\ell_Q(w_k) \approx \frac{1}{2}(z(\tilde{m}) - Xw_k)^T \ell''(\tilde{m})(z(\tilde{m}) - Xw_k) + C(\tilde{m}, \tilde{w}_k),$$

where $\tilde{m} = X\tilde{w}_k$, $z(\tilde{m}) = \tilde{m} - \ell''(\tilde{m})^{-1}\ell'(\tilde{m})$, and $C(\tilde{m}, \tilde{w}_k)$ consist of the rest of the terms that do not depend on w_k.

As stated in the first stage of feature extraction, the longitudinal profile of a time-varying biomarker can be summarized as a set of spline coefficients, which are further treated as a grouped feature. To accommodate this grouping structure and conduct group-feature selection, the group LASSO penalty (Yuan and Lin, 2006) has been incorporated and the objective function is defined as:

$$L_\tau(w_k) = \frac{1}{2N}\left(z(\tilde{m}) - Xw_k\right)^T \ell''(\tilde{m})(z(\tilde{m}) - Xw_k) + \tau \sum_{g=1}^{G} \sqrt{p_g}||w_{k_g}||_2, \qquad (3.3.3)$$

where w_{k_g} is a vector of regression coefficients corresponding to the g^{th} groupded feature, p_g denotes the number of spline coefficients within g^{th} group which is used to adjust for the varying group sizes, G is the total number of groups, and $||\cdot||_2$ represents the L_2 norm. It is worth noting that the tuning parameter τ is usually determined by a $\mathcal{D}$-fold cross-validation which involves the bias-variance trade-off. Specifically, a small $\mathcal{D}$ typically results in large bias but small variance, whereas a large $\mathcal{D}$ results in small bias but large variance. Considering the sample size of the minority class, a 5-fold cross-validation is often used to choose the optimal value for τ.

To efficiently solve the penalized reweighted least squares in Equation (3.3.3), the block coordinate descent algorithm is employed for optimization. Each vector of grouped coefficients, i.e., w_{k_g}, is estimated iteratively on the block level. Specifically, for the q^{th} grouped feature, the first derivative of $L_\tau(w_k)$ with respect to w_{k_q} is computed as:

$$\frac{\partial L_\tau(w_k)}{\partial w_{k_q}} = -\frac{1}{N} X_q^T \ell''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq q} X_g w_{k_g} - X_q w_{k_q} \right) + \tau \sqrt{p_q} s_q, \tag{3.3.4}$$

where X_g and X_q are the data matrices corresponding to the g^{th} and q^{th} grouped features, respectively, p_q is the group size of q^{th} grouped feature, and

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

Hence, $\hat{w}_{k_q}$ can be easily obtained by setting Equation (3.3.4) to zero. To be more precise, if $w_{k_q} = \mathbf{0}$, the following inequality holds:

$$\left\| \frac{1}{N} X_q^T \ell''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq q} X_g w_{k_g} \right) \right\|_2 \leqslant \tau \sqrt{p_q}, \tag{3.3.5}$$

and if $w_{k_q} \neq \mathbf{0}$, it has a closed-form solution as follows:

$$\hat{w}_{k_q} = \left[\frac{1}{N} X_q^T \ell''(\tilde{m}) X_q + \frac{\tau \sqrt{p_q}}{||w_{k_q}||_2} \cdot I \right]^{-1} \cdot \left[\frac{1}{N} X_q^T \ell''(\tilde{m}) \left(z(\tilde{m}) - \sum_{g \neq q} X_g w_{k_g} \right) \right]. \tag{3.3.6}$$

Algorithm 2 Inner loop for the estimation of w_k

```
Step 1. Initialize $\tilde{w}_k$, and compute $\tilde{m}$, $\ell^{'}(\tilde{m})$, $\ell^{''}(\tilde{m})$, and $z(\tilde{m})$.
Step 2. For $q = 1, ..., G$, if Equation (3.3.5) holds, $\hat{w}_{k_q}$ is set to $\mathbf{0}$; otherwise, $\hat{w}_{k_q}$ is updated
        using Equation (3.3.6).
Step 3. Set $\tilde{w}_k = \hat{w}_k$, and compute $\tilde{m}$, $\ell^{'}(\tilde{m})$, $\ell^{''}(\tilde{m})$, and $z(\tilde{m})$.
Step 4. Repeat steps 2 - 3 until convergence.
```

This optimization procedure requires two loops to search for the optimal solution path: (1) for the outer loop, a partial quadratic approximation is applied to obtain $\ell_Q(w_k)$ for each $k \in \{1, ..., C\}$, and (2) for the inner loop, $\hat{w}_k$ can be estimated efficiently by using the Algorithm 2. To speed up the computation in the inner loop, a strategy called *active-set* convergence (Krishnapuram et al., 2005; Meier et al., 2008; Friedman et al., 2010) has been adopted. Specifically, an *active-set* is generated after the first cycle through G groups. The remaining iterations are then restricted to this *active-set* which is updated after each following cycle. The entire process stops when the *active-set* does not change.

The primary goal of the proposed optimization framework is to estimate the coefficients vectors $\{w_c\}_{c=1,...,C}$ which can be used to make predictions for the belonging class of unseen data. Hence, given any new instance with a p-dimensional vector x_{new}, it will be classified into the class which outputs the largest score $\hat{w}_c^T x_{new}$. It is worth noting that this algorithm is still applicable to binary case by simply setting C to two, despite that it is mainly developed for multi-class (C ≥3) scenario.

3.4 Performance metrics for classification evaluation

A number of metrics have been developed and widely used to evaluate the performance of a binary classifier. For example, accuracy and error rate are two popular measures for balanced data. Additionally, to accommodate the imbalanced structure, some other metrics have been employed, such as precision, recall, F1 score and G-mean. However, these measures are not readily used for multi-class classification. In our study, we adopt the multi-class variants of those aforementioned metrics (Opitz and Burst, 2019; Grandini et al., 2020; Tanha et al., 2020) for comparison purposes throughout all numerical analyses. Among them, the most popular ones are macro-precision and macro-recall that are simply the arithmetic mean of precision and recall for each of C classes, respectively:

$$\text{macro-precision} = \frac{\sum_{c=1}^{C} \text{precision}_c}{C},$$

$$\text{macro-recall} = \frac{\sum_{c=1}^{C} \text{recall}_c}{C},$$

where precision = TP /(TP + FP) and recall = TP / (TP + FN). Note that TP, FP and FN represent true positive, false positive and false negative, respectively.

Then the macro-F1 score is computed as the harmonic mean of macro-precision and macro-recall:

$$\text{macro-F1 score} = \frac{2 \times \text{macro-precision} \times \text{macro-recall}}{\text{macro-precision}^{-1} + \text{macro-recall}^{-1}}.$$

Lastly, the G-mean can be computed as:

$$\text{G-mean} = \left(\prod_{c=1}^{C} \text{recall}_c \right)^{\frac{1}{C}}.$$

It should be noted that all these metrics proposed for multi-class classification have a range from 0 to 1, and a higher value typically indicates a better performance.

3.5 Simulation Study

In this section, extensive simulations have been conducted to evaluate the performance of the proposed multi-class framework. Typically, we consider two settings to generate data: (i) a low-dimensional setting with $n > p$ and (ii) a high-dimensional setting with $n < p$. For each setting, we further assess the classification performance under two scenarios, i.e., the labels across classes are either imbalanced or balanced. It is assumed through all simulations that the longitudinal observations of each subject are measured at seven discrete time points (i.e., $t \in \{0, 1, 2, 3, 4, 5, 6\}$ with $t = 0$ representing the baseline).

Basically, the data-generating scheme is a two-stage procedure. Longitudinal predictors are simulated for each of the three classes (C_1, C_2 and C_3) separately in the first stage:

$$\begin{aligned} U_{ej}^{C_1}(t) &= \mu_j^{C_1}(t) + b_{ej} + \varepsilon_{ej}(t),\ t \in \{0, 1, ..., 6\}, \\ U_{rj}^{C_2}(t) &= \mu_j^{C_2}(t) + b_{rj} + \varepsilon_{rj}(t),\ t \in \{0, 1, ..., 6\}, \\ U_{sj}^{C_3}(t) &= \mu_j^{C_3}(t) + b_{sj} + \varepsilon_{sj}(t),\ t \in \{0, 1, ..., 6\}, \end{aligned}$$

where $\mu_j^{C_1}(t)$, $\mu_j^{C_2}(t)$ and $\mu_j^{C_3}(t)$ are the mean functions of j^{th} longitudinal predictor for each class. The subject-specific random effects b_{ej}, b_{rj} and b_{sj} correspond to the e^{th} C_1 subject, r^{th} C_2 subject and s^{th} C_3 subject, respectively, and are all generated from $N(0, 1.5)$. The random errors $\varepsilon_{ej}(t)$, $\varepsilon_{rj}(t)$ and $\varepsilon_{sj}(t)$ are generated from $N(0, 1)$. In the second stage, each longitudinal profile is transformed into a set of spline coefficients using the techniques of natural cubic spline. Then these sets of spline coefficients are combined and treated as features which are further used in subsequent classification procedure. It is worth noting that the number of inner knots is set to two during the process of feature extraction from longitudinal predictors for simplicity.

In our analysis, regardless of being in low- or high-dimensional settings, the first 25 longitudinal predictors are considered significant for simplicity throughout all simulations and the corresponding mean functions are generated separately for each of the three classes. For class C_1, the mean function $\mu_j^{C_1}(t)$ is assumed to be constant, i.e. $\mu_j^{C_1} = (\phi_j^{C_1}, \phi_j^{C_1}, ..., \phi_j^{C_1})^T_{1\times 7}$, $j \in \{1, ..., 25\}$, where $\phi_j^{C_1}$ is sampled from a truncated normal distribution (TN):

$$\phi_j^{C_1} \sim TN(0, 1),\ \phi_j^{C_1} \in [-1, 1].$$

For class C_2, let $\mu_j^{C_2} = \left\{\gamma_{0j}^{C_2} + \gamma_{1j}^{C_2}\ t + \gamma_{2j}^{C_2}\ t^2\right\}_{t=0,1,...,6}$, $j \in \{1, ..., 25\}$. The coefficients $\{\gamma_{\rho j}^{C2}\}_{\rho=0,1,2}$ that are associated with the j^{th} longitudinal predictor are randomly chosen from truncated normal distributions:

$$\begin{aligned}
\gamma_{0j}^{C_2} &\sim TN(0,1) \quad , \quad \gamma_{0j}^{C_2} \in [-1,1],\\
\gamma_{1j}^{C_2} &\sim TN(0,1) \quad , \quad \gamma_{1j}^{C_2} \in [-0.1,0.1],\\
\gamma_{2j}^{C_2} &\sim TN(0,1) \quad , \quad \gamma_{2j}^{C_2} \in [-0.01,0.01].
\end{aligned}$$

For class C_3, let $\mu_j^{C_3} = \{\gamma_{0j}^{C_3} + \gamma_{1j}^{C_3}\ t + \gamma_{2j}^{C_3}\ t^3\}_{t=0,1,\ldots,6}$, $j \in \{1, ..., 25\}$, and the corresponding coefficients $\{\gamma_{\rho j}^{C3}\}_{\rho=0,1,2}$ are generated from the truncated normal distributions:

$$\begin{aligned}
\gamma_{0j}^{C_3} &\sim TN(0,1) \quad , \quad \gamma_{0j}^{C_3} \in [-1,1],\\
\gamma_{1j}^{C_3} &\sim TN(0,1) \quad , \quad \gamma_{1j}^{C_3} \in [-0.1,0.2],\\
\gamma_{2j}^{C_3} &\sim TN(0,1) \quad , \quad \gamma_{2j}^{C_3} \in [0,0.001].
\end{aligned}$$

As for the remaining insignificant predictors in each setting, we assume the mean functions to be the same for all three classes, that is, $\mu_j^{C_1} = \mu_j^{C_2} = \mu_j^{C_3} = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)^T$.

To evaluate the performance of the proposed method, high and low dimensional settings are examined separately by assigning different numbers of longitudinal predictors. Specifically, we use values $\{250, 500\}$ for high-dimensional setting and $\{50, 100\}$ for low-dimensional setting. For each setting, the number of subjects in each class is adjusted to generate imbalanced or balanced data structure. The sample sizes of three different classes for both training and test data are set to $(150, 90, 30)$ for imbalanced scenario and set to $(100, 100, 100)$ for balanced scenario. For comparison purposes, another three popular approaches, i.e., multi-class support vector machines (SVM) (Fan et al., 2005), linear discriminant analysis (LDA) (Ripley, 2007), and local mean-based k-nearest centroid neighbor (LMKNCN) (Gou et al., 2012), have also been performed. In each scenario, 500 Monte Carlo replicates are simulated.

3.5.1 High-dimensional setting

(i) *Imbalanced data:* The simulation results are provided in Table 3.1. It seems that SVM tends to classify subjects into the most weighted majority class, thus failing to generate macro precision and F1 score and yielding same G-mean, macro recall and overall accuracy. For the state-of-the-art LMKNCN, severe overfitting leads to terrible performances on test data. In addition, LDA has the same severe misclassification problem as SVM and performs poorly when the dimensionality of data is ultra high ($p = 1000$). In contrast, our proposed approach is capable of dealing with the imbalanced data structure and overfitting issue, and outperforms the other three methods in terms of all five metrics on test data.

Table 3.1: Results (S.E.) of multi-class ($C = 3$) classification for imbalanced data under high-dimensional settings based on 500 Monte Carlo replicates

Number of longitudinal features				SVM	LDA	LMKNCN	Proposed Method
500	$p = 1000$	Training	G-Mean	.000(.000)	.591(.040)	.999(.001)	.858(.035)
			Macro Recall	.333(.001)	.614(.031)	.999(.001)	.861(.035)
			Macro Precision	–†	.680(.038)	.999(.001)	.784(.040)
			Macro F1 Score	–†	.645(.032)	.999(.001)	.821(.037)
			Overall accuracy	.556(.001)	.699(.023)	.999(.001)	.831(.037)
		Test	G-Mean	.000(.000)	.000(.001)	.438(.077)	.683(.068)
			Macro Recall	.333(.001)	.384(.037)	.420(.053)	.702(.075)
			Macro Precision	–†	–†	.463(.118)	.666(.064)
			Macro F1 Score	–†	–†	.202(.185)	.688(.089)
			Overall accuracy	.556(.001)	.578(.041)	.570(.048)	.731(.059)
250	$p = 500$	Training	G-Mean	.000(.000)	.825(.032)	.999(.001)	.859(.044)
			Macro Recall	.333(.001)	.828(.031)	.999(.001)	.862(.043)
			Macro Precision	–†	.848(.029)	.999(.001)	.789(.050)
			Macro F1 Score	–†	.838(.028)	.999(.001)	.824(.046)
			Overall accuracy	.556(.001)	.851(.024)	.999(.001)	.832(.046)
		Test	G-Mean	.000(.000)	.466(.077)	.466(.084)	.692(.074)
			Macro Recall	.333(.001)	.452(.063)	.444(.058)	.713(.081)
			Macro Precision	–†	.484(.100)	.498(.128)	.674(.070)
			Macro F1 Score	–†	.281(.190)	.227(.190)	.702(.095)
			Overall accuracy	.556(.001)	.589(.060)	.596(.057)	.732(.066)

LMKNCN: local mean-based k-nearest centroid neighbor; p: number of transformed features in the simulated data; †: fail to generate.

(ii) *Balanced data:* As shown in Table 3.2, unlike the imbalanced scenario under high-dimensional setting, SVM successfully generates all five metrics, and its performances on test data are much better than LMKNCN and LDA. Under this balanced data structure, LMKNCN is still experiencing overfitting issue but the performances across different metrics are relatively close to each other compared to the particularly low macro F1 score under imbalanced structure. Moreover, our proposed approach still outperforms the other three methods on test data and noticeable improvements have also been observed compared to the imbalanced scenario.

Table 3.2: Results (S.E.) of multi-class ($C = 3$) classification for balanced data under high-dimensional settings based on 500 Monte Carlo replicates

Number of longitudinal features				SVM	LDA	LMKNCN	Proposed Method
500	$p = 1000$	Training	G-Mean	.981(.011)	.651(.023)	.999(.001)	.866(.040)
			Macro Recall	.981(.011)	.652(.023)	.999(.001)	.867(.040)
			Macro Precision	.982(.010)	.653(.023)	.999(.001)	.868(.040)
			Macro F1 Score	.982(.010)	.652(.023)	.999(.001)	.867(.040)
			Overall accuracy	.981(.011)	.652(.023)	.999(.001)	.867(.040)
		Test	G-Mean	.621(.066)	.459(.061)	.459(.058)	.755(.063)
			Macro Recall	.615(.066)	.459(.061)	.457(.057)	.753(.063)
			Macro Precision	.628(.068)	.460(.062)	.460(.059)	.757(.063)
			Macro F1 Score	.602(.072)	.451(.062)	.450(.058)	.749(.063)
			Overall accuracy	.615(.066)	.459(.061)	.457(.057)	.753(.063)
250	$p = 500$	Training	G-Mean	.943(.018)	.851(.023)	.999(.001)	.877(.039)
			Macro Recall	.944(.018)	.851(.023)	.999(.001)	.878(.039)
			Macro Precision	.946(.016)	.852(.023)	.999(.001)	.878(.039)
			Macro F1 Score	.945(.017)	.852(.023)	.999(.001)	.878(.039)
			Overall accuracy	.944(.018)	.851(.023)	.999(.001)	.878(.039)
		Test	G-Mean	.691(.059)	.517(.061)	.505(.054)	.766(.054)
			Macro Recall	.685(.060)	.516(.060)	.503(.054)	.764(.054)
			Macro Precision	.697(.060)	.519(.062)	.506(.055)	.769(.054)
			Macro F1 Score	.675(.670)	.509(.061)	.497(.055)	.761(.055)
			Overall accuracy	.685(.060)	.516(.060)	.503(.054)	.764(.054)

LMKNCN: local mean-based k-nearest centroid neighbor; p: number of transformed features in the simulated data.

3.5.2 Low-dimensional setting

(i) *Imbalanced data:* As shown in Table 3.3, overfitting still occurs in both LDA and LMKNCN even under the low-dimensional setting. The imbalanced structure leads to a relatively low macro F1 score in LMKNCN. Besides that, the severe misclassification has not been improved in SVM under this scenario. The proposed method still outperforms the other three classifiers on test data and this low-dimensional setting results in observable improvements across all five metrics compared to the performances of our method in high-dimensional imbalanced setting.

Table 3.3: Results (S.E.) of multi-class ($C = 3$) classification for imbalanced data under low-dimensional settings based on 500 Monte Carlo replicates

Number of longitudinal features				SVM	LDA	LMKNCN	Proposed Method
100	$p = 200$	Training	G-Mean	.000(.000)	.999(.001)	.999(.001)	.887(.043)
			Macro Recall	.333(.001)	.999(.001)	.999(.001)	.889(.043)
			Macro Precision	–†	.999(.001)	.999(.001)	.822(.050)
			Macro F1 Score	–†	.999(.001)	.999(.001)	.854(.046)
			Overall accuracy	.556(.001)	.999(.001)	.999(.001)	.863(.044)
		Test	G-Mean	.000(.000)	.434(.063)	.550(.084)	.717(.065)
			Macro Recall	.333(.001)	.442(.072)	.508(.062)	.736(.072)
			Macro Precision	–†	.428(.057)	.607(.129)	.700(.063)
			Macro F1 Score	–†	.424(.080)	.347(.182)	.724(.083)
			Overall accuracy	.556(.001)	.459(.068)	.658(.059)	.760(.056)
50	$p = 100$	Training	G-Mean	.000(.000)	.981(.013)	.999(.001)	.897(.047)
			Macro Recall	.333(.001)	.982(.013)	.999(.001)	.899(.046)
			Macro Precision	–†	.983(.012)	.999(.001)	.841(.062)
			Macro F1 Score	–†	.982(.012)	.999(.001)	.869(.054)
			Overall accuracy	.556(.001)	.982(.012)	.999(.001)	.876(.050)
		Test	G-Mean	.000(.000)	.607(.076)	.618(.090)	.743(.059)
			Macro Recall	.333(.001)	.612(.079)	.572(.073)	.765(.063)
			Macro Precision	–†	.603(.077)	.678(.124)	.724(.060)
			Macro F1 Score	–†	.584(.110)	.450(.183)	.757(.069)
			Overall accuracy	.556(.001)	.672(.064)	.709(.057)	.777(.053)

LMKNCN: local mean-based k-nearest centroid neighbor; p: number of transformed features in the simulated data; †: fail to generate.

(ii) *Balanced data:* Under this low-dimensional balanced structure, the performances of the proposed method and SVM are comparable whereas LDA and LMKNCN are still suffering from the overfitting issue. The simulation results based on 500 Monte Carlo replicates are given in Table 3.4.

Table 3.4: Results (S.E.) of multi-class ($C = 3$) classification for balanced data under low-dimensional settings based on 500 Monte Carlo replicates

Number of longitudinal features				SVM	LDA	LMKNCN	Proposed Method
100	$p = 200$	Training	G-Mean	.881(.030)	.999(.001)	.999(.001)	.876(.039)
			Macro Recall	.883(.029)	.999(.001)	.999(.001)	.876(.039)
			Macro Precision	.888(.027)	.999(.001)	.999(.001)	.877(.039)
			Macro F1 Score	.885(.028)	.999(.001)	.999(.001)	.877(.039)
			Overall accuracy	.883(.029)	.999(.001)	.999(.001)	.876(.039)
		Test	G-Mean	.740(.061)	.412(.058)	.564(.061)	.765(.055)
			Macro Recall	.734(.062)	.410(.056)	.563(.060)	.762(.055)
			Macro Precision	.745(.061)	.414(.060)	.566(.061)	.767(.055)
			Macro F1 Score	.726(.067)	.392(.064)	.557(.062)	.758(.056)
			Overall accuracy	.734(.062)	.410(.056)	.563(.060)	.762(.055)
50	$p = 100$	Training	G-Mean	.836(.051)	.965(.020)	.999(.001)	.863(.058)
			Macro Recall	.840(.048)	.965(.020)	.999(.001)	.864(.058)
			Macro Precision	.849(.043)	.965(.020)	.999(.001)	.865(.057)
			Macro F1 Score	.844(.045)	.965(.020)	.999(.001)	.845(.057)
			Overall accuracy	.840(.048)	.965(.020)	.999(.001)	.864(.058)
		Test	G-Mean	.763(.070)	.651(.074)	.643(.081)	.755(.067)
			Macro Recall	.755(.071)	.649(.074)	.640(.080)	.753(.067)
			Macro Precision	.770(.070)	.653(.075)	.645(.081)	.758(.067)
			Macro F1 Score	.746(.077)	.643(.075)	.636(.081)	.750(.068)
			Overall accuracy	.755(.071)	.649(.074)	.640(.080)	.753(.067)

LMKNCN: local mean-based k-nearest centroid neighbor; p: number of transformed features in the simulated data.

3.6 Real applications

3.6.1 Human walking data

Data used in this analysis are from an exploratory study (Chang et al., 2020) which aimed at differentiating gender by analysis of walking patterns. Sixty-one college-aged subjects (25 females & 36 males) were asked to walk in a ten-foot path, turn around, and then walk back to the original location, which is referred to as a complete gait cycle. Two Microsoft Kinect cameras were used to create the 3D kinematic view of 25 joints from the human body and record the walking pattern, i.e., the X, Y, and Z coordinates of all joints, from the front and side view of each subject. More information regarding how the (x, y, z) coordinate space is constructed and measured is available at https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/en-us-events-fs2011-jancke_kinect_programming.pdf. Figure 3.1 depicts the time-series gait profiles of five randomly selected subjects. As indicated by this unpublished exploratory study, the performances of models with side-view data are almost identical to those of the use of all data. Therefore, only the side-view data with all three coordinates are included in our study for person recognition. In addition, we also observe that different subjects may have varying numbers of recorded cycles, which leads to another complication of data imbalance among multiple classes.

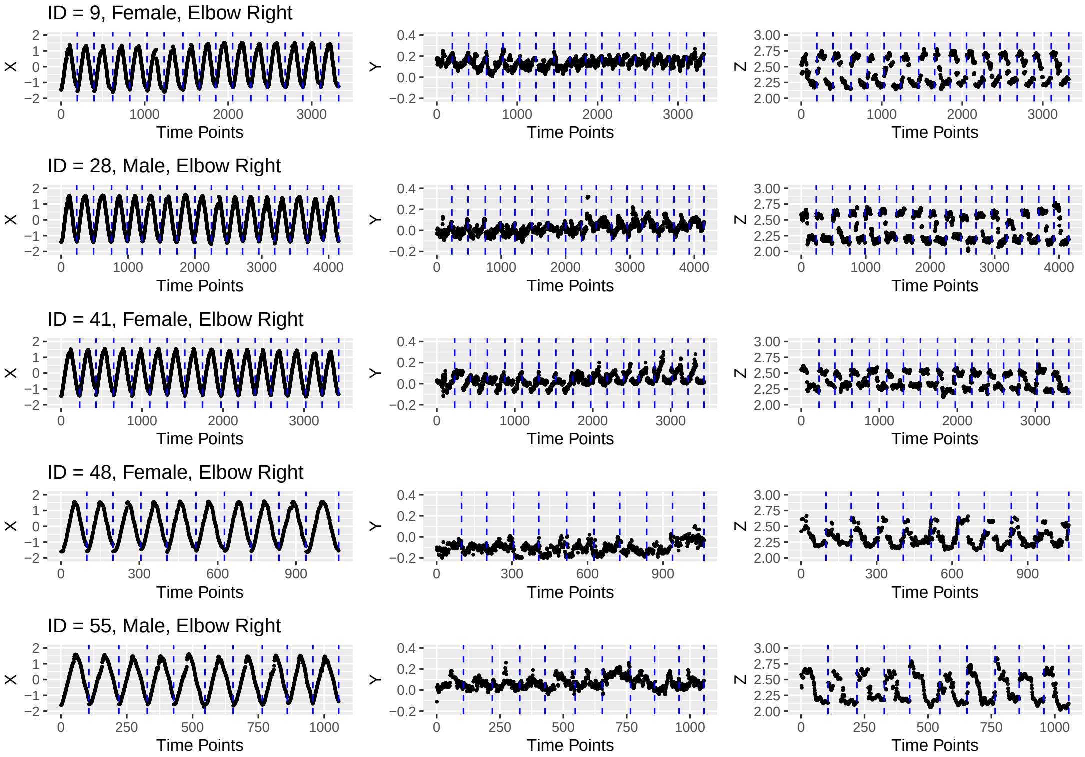

Figure 3.1: Time-series gait profiles of five randomly selected subjects, i.e., ID = 9, 28, 41, 48 and 55, in terms of X, Y and Z coordinates across different trials.

To deal with the dense longitudinal profiles of three coordinates from all 25 joints in the gait data, the techniques of natural cubic spline are applied to obtain the corresponding spline coefficients which are then used as cross-sectional features in subsequent classification procedure. All these extracted coefficients are standardized with zero mean and unit variance. For the model evaluation, one gait cycle is randomly selected from each of the 61 subjects as test data, and the rest of cycles are used as training data. It should be noted that only six gait cycles are available for some subjects, which is why we only use one cycle from each subject in the test data, thus ensuring sufficient data for model training. The optimal tuning parameter is determined by cross-validation in the model training stage. Because only one gait cycle is used in the test data for each subject, it is impossible to calculate those metrics employed earlier, such as G-mean and macro recall. For comparison purpose, the overall accuracy, i.e., the proportion of correct identification out of 61 subjects, is only considered as the performance metric for the test data. As shown in Table 3.5, the performances of all four methods on training data are comparable. However, a significant drop is observed in the overall accuracy for both SVM and LMKNCN in the test data, whereas our proposed method and LDA retain exceptional and similar performances. It is worth mentioning that the performance of LDA depends heavily on the estimation of population covariance matrix and mean vectors, which becomes challenging in high-dimensional settings (Sifaou et al., 2020). In this gait dataset, the mean differences in the overall walking patterns among subjects are noticeable and the dimension of transformed data is relatively low, which is why LDA achieves such a high overall accuracy for test data.

3.6.2 Alzheimer's disease data

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a multi-site longitudinal study whose primary goal has been to test whether a combination of different types of data, such as serial magnetic resonance (MRI), positron tomography (PET), biological and genetic biomarkers, and clinical assessments, can be used to track the progression and early predict the conversion of Alzheimer's disease (AD). In the ADNI database (adni.loni.usc.edu), each

Table 3.5: Results (S.E.) of gait-based person recognition using 61 subjects (i.e. classes) based on 500 Monte Carlo replicates

		SVM	LDA	LMKNCN	Proposed Method
Training	G-Mean	.956(.010)	.999(.001)	.989(.004)	.999(.001)
	Macro Recall	.964(.006)	.999(.001)	.990(.003)	.999(.001)
	Macro Precision	.980(.004)	.999(.001)	.992(.003)	.999(.001)
	Macro F1 Score	.972(.005)	.999(.001)	.991(.003)	.999(.001)
	Overall accuracy	.968(.005)	.999(.001)	.990(.003)	.999(.001)
Test	Overall accuracy†	.727(.048)	.913(.033)	.788(.047)	.907(.036)

LMKNCN: local mean-based k-nearest centroid neighbor: †: calculated based on only one observation.

participant has a longitudinal profile with measurements conducted repeatedly at a six-month interval. During each clinical visit, participants generally undergo a series of assessments and are labeled with: cognitively normal (CN), mild cognitive impairment (MCI) or AD.

In our study, we mainly focus on subjects who are diagnosed as CN at baseline and investigate the conversion process over time. In other words, it is of our interest to develop a prognostic model which can be used for the early prediction of AD among CN subjects. Meanwhile, both the stable and MCI-converted CN subjects can be monitored dynamically and the prediction of whether the conversion will occur (i.e. CN converts to MCI or MCI converts to AD) is updated whenever additional data become available. In this case, we perform classification for three classes. Figure 3.2 shows the possible conversion processes of baseline normal subjects and how training data are determined to achieve the goals mentioned above. We select 319 subjects (CN: 237, MCI: 52 and AD: 30) who are diagnosed as normal at baseline and each of them has at least three clinical visits. The demographic information is summarised in Table 3.6. Besides that, the longitudinal data in the ADNI database are irregularly and sparsely collected. More specifically, the assessments are generally conducted at discrete time points which vary across participants, thus leading to different numbers of visits among subjects. The distribution of the number of visits is given in Table 3.7.

In this analysis, different types of biomarkers and scores have been included in our model. Due to the high association between the brain abnormalities detected by MRI and the pro-

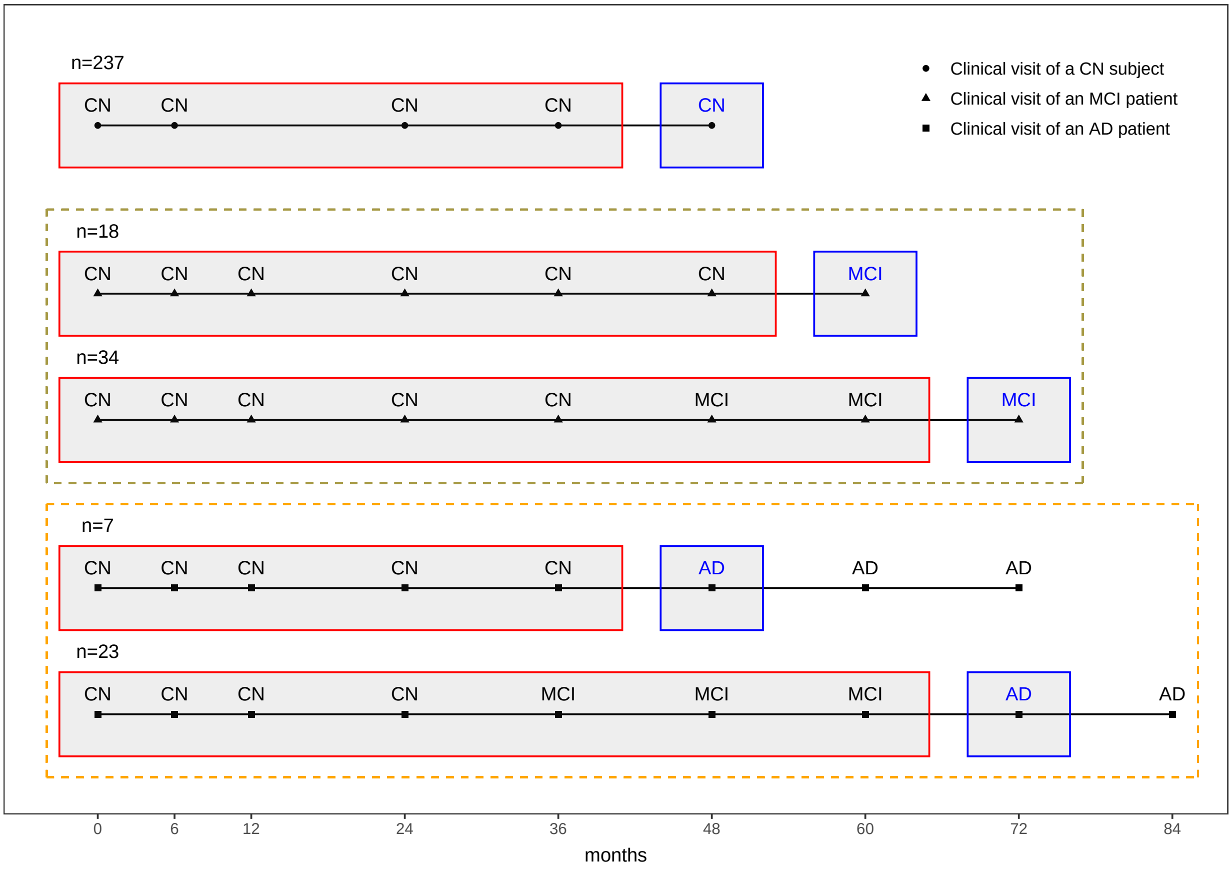

Figure 3.2: Clinical diagnosis of a CN subject, an MCI patient or an AD patient over time. The dashed rectangle shows different conversion processes for MCI and AD patients, respectively. The red box represents the data used for model training. The blue box represents the final diagnosis used as the membership outcome.

gression of AD (Frisoni et al., 2010; Zhang et al., 2016; Gavidia-Bovadilla et al., 2017; Long et al., 2017; Huang et al., 2017), we mainly focus on longitudinal biomarkers extracted from the MRI modality. With the use of Freesurfer (v6.0.0, https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki), a total number of 319 biomarkers are generated from various regions of interest (ROI) in the brain. Moreover, another five popular cognitive and functional scores are also used (Li et al., 2017; Lin et al., 2020): Functional Assessment Questionnaire (FAQ), Mini Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-Cognitive 13 items (ADAS-Cog 13), Rey Auditory Verbal Learning Tests (RAVLT immediate score and RAVLT learning score). In addition to these longitudinal features, we also consider several baseline demographic and genetic variables in the model: gender, age, and apolipoprotein E

Table 3.6: Demographic characteristics of selected subjects

		Age (years)		Gender (%)	
Group	n	Mean	Std. Dev	Male	Female
CN	237	74.5	5.6	52.7	47.3
MCI	52	76.3	5.4	59.6	40.4
AD	30	75.4	3.9	40.0	60.0

Table 3.7: Distribution of number of visits

	Number of subjects		
Visits	CN	MCI	AD
3	68	13	2
4	100	11	4
5	13	12	3
6	10	3	5
7	13	3	2
8	14	7	4
9	10	3	7
10	9	0	3
Total	237	52	30

allele $\varepsilon 4$ (APOE4), which might be predictive of the conversion of MCI and AD. To illustrate the irregular and sparse structure of the ADNI data, we use the ADAS-Cog 13 as an example and present the longitudinal trajectories of CN subjects, MCI and AD patients in Figure 3.3. The differences in the overall trends between three groups are easily observed, indicating the potential of ADAS-Cog 13 in discriminating three distinct stages in the progression of AD.

For the model evaluation, 30% of subjects from each class are randomly selected as test data, and the remaining subjects are used as training data. All the transformed features from original longitudinal predictors by using natural cubic spline are standardized with zero mean and unit variance. The tuning parameter τ in the penalty term is determined by a 5-fold cross-validation. The results over 500 Monte Carlo replicates are provided in Table 3.8. It seems that overfitting occurs to both LDA and LMKNCN and their corresponding performances on test data are relatively poor. SVM still fails to output the macro precision

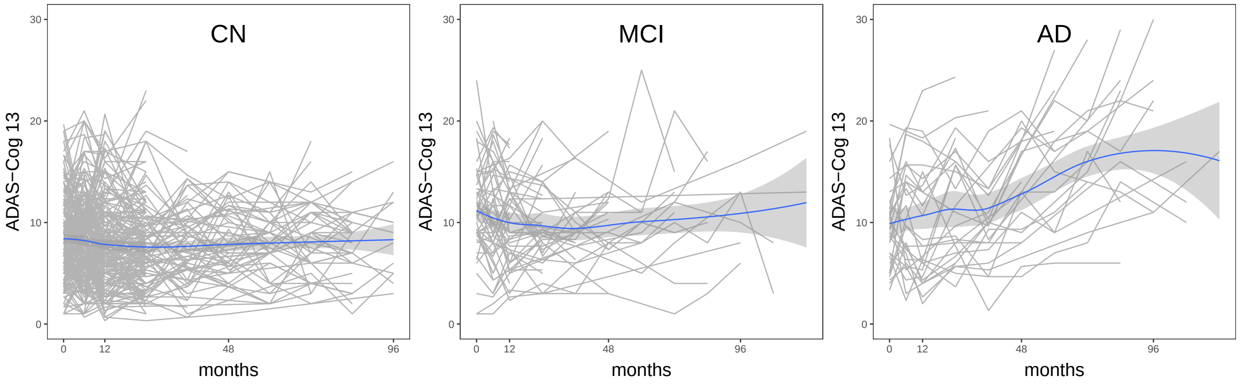

Figure 3.3: Longitudinal trajectories of ADAS-Cog 13 for CN subjects, MCI and AD patients.

and F1 score due to the imbalanced structure (CN: 237, MCI: 52 and AD: 30). However, our proposed method is capable of dealing with the class imbalance and outperforms the other three approaches under the high-dimensional setting, especially in terms of G-mean, macro-recall and F1 score. As shown in Table 3.8, the proposed multi-class framework can still achieve approximately 65% G-mean and 67% macro-recall, respectively, even for such a complex imbalanced data, whereas LDA and LMKNCN only achieve 40+% G-mean and 53% macro-recall.

Table 3.8: Results (S.E.) of early detection of Alzheimer's disease and prediction of conversion of the other two stages using ADNI data based on 500 Monte Carlo replicates

		SVM	LDA	LMKNCN	Proposed Method
Training	G-Mean	.000(.000)	.951(.015)	.999(.003)	.727(.043)
	Macro Recall	.333(.001)	.953(.014)	.999(.003)	.738(.038)
	Macro Precision	–†	.975(.011)	.999(.001)	.658(.043)
	Macro F1 Score	–†	.964(.011)	.999(.002)	.695(.040)
	Overall accuracy	.740(.001)	.972(.009)	.999(.001)	.758(.027)
Test	G-Mean	.000(.000)	.464(.097)	.418(.084)	.650(.045)
	Macro Recall	.333(.001)	.538(.064)	.528(.053)	.668(.042)
	Macro Precision	–†	.593(.072)	.612(.080)	.614(.044)
	Macro F1 Score	–†	.563(.061)	.565(.057)	.639(.037)
	Overall accuracy	.735(.001)	.714(.038)	.752(.030)	.723(.036)

LMKNCN: local mean-based k-nearest centroid neighbor; †: fail to generate.

3.6.3 Leukemia data

The leukemia data were first described by Golub et al. (1999) and have been frequently used in many microarray studies (Dudoit et al., 2002; Dettling and Bühlmann, 2003; Huang and Zheng, 2006). The primary goal is to determine the leukemia phenotypes using gene expression measurements which were obtained from Affymetrix high-density oligonucleotide microarrays. The complete dataset is available at: https://hastie.su.domains/CASI_files/ DATA/leukemia.html. Generally, there are two main types of leukemia, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). If a subject is diagnosed as ALL, further assessments need to be conducted to determine whether it is T-cell or B-cell. Therefore, three types of leukemia are considered in this study: AML, ALL-T and ALL-B. According to the preprocessing procedure described by Dudoit et al. (2002), the log-transformation and standardization have been applied to the original data. For our analysis, a total number of 72 patients (ALL-B: 38, ALL-T: 9 and AML:25) are included, with each containing 7129 human genes.

In implementing the proposed algorithm, the multi-class classification framework has been slightly modified to accommodate the cross-sectional leukemia dataset. To be more precise, the feature extraction step via natural cubic spline is skipped because no longitudinal profiles are present. For the model evaluation, this dataset is randomly divided into training and test subsets, each containing 70% and 30% of leukemia patients, respectively. The tuning parameter is determined by a 3-fold cross-validation. As shown in Table 3.9, the proposed method outperforms the other three competitive approaches regardless of which metric is employed. SVM still fails to generate the macro precision and F1 score because it tends to classify all subjects as ‘ALL-B’. Unlike its performance on Alzheimer’s disease data, LDA performs poorly even on the training subset due to the ultra high-dimensionality of this leukemia dataset. For the state-of-the-art LMKNCN, its performance on the test data is also inferior to that of our proposed method.

Table 3.9: Results (S.E.) of tumor types discrimination using Leukemia data based on 500 Monte Carlo replicates

		SVM	LDA	LMKNCN	Proposed Method
Training	G-Mean	.000(.000)	.762(.052)	.999(.001)	.999(.002)
	Macro Recall	.333(.001)	.795(.041)	.999(.001)	.999(.002)
	Macro Precision	–†	.926(.034)	.999(.001)	.999(.006)
	Macro F1 Score	–†	.854(.034)	.999(.001)	.999(.004)
	Overall accuracy	.531(.001)	.869(.024)	.999(.001)	.999(.004)
Test	G-Mean	.000(.000)	.764(.115)	.820(.099)	.928(.057)
	Macro Recall	.333(.001)	.778(.094)	.838(.084)	.932(.052)
	Macro Precision	–†	.829(.039)	.922(.036)	.931(.048)
	Macro F1 Score	–†	.802(.069)	.876(.059)	.931(.046)
	Overall accuracy	.522(.001)	.836(.068)	.873(.057)	.924(.048)

LMKNCN: local mean-based *k*-nearest centroid neighbor; †: fail to generate.

3.7 Discussion

In this work, we have developed a novel classification framework for multi-class data with a complex structure, such as functional/longitudinal measurements, high-dimensionality and class imbalance. With the use of natural cubic spline, the dense longitudinal/functional data can be efficiently characterized by a set of spline coefficients with a substantial dimensional reduction. A novel exponential loss function coupled with group LASSO penalty is then constructed to cast the multi-class classification task as one single optimization problem. Additionally, a simple weight-adjusted margin can be easily incorporated into the proposed loss function to address the issue of class imbalance. The proposed approach does not require any distributional assumptions on biomarkers and is capable of performing feature selection for all classes individually.

It is worth mentioning that natural cubic spline is not the only technique for extracting features from longitudinal/functional data. When only few measurements, such as one or two, are collected for certain subjects, it may not be feasible to transform those functional profiles using natural cubic spline. Instead, functional principal component analysis (FPCA) can be used to handle this issue. To accommodate the irregular and sparse longitudinal data, Yao et al. (2005) proposed the Principal Components Analysis through Conditional Expectation (PACE) algorithm which uses pooled data to estimate the so-called functional principal component (FPC) scores regardless of how many measurements exist for each subject. Another complication with the use of natural cubic spline is that it generally requires a pre-specified number of inner knots to apply natural cubic spline. A simulation study has been conducted to investigate how the number of inner knots will impact the classification performance. The results (not shown here) indicate that two inner knots will be sufficient to characterize the over-time underlying associations in functional/longitudinal profiles.

Chapter 4

Conclusion and Discussion

In this dissertation, we propose two novel frameworks for the binary and multi-class classification of complex imbalanced data. In the literature, existing classifiers generally assumed a balanced class distribution which is often not true in biomedical research, especially in rare disease screening and early diagnosis studies. Considering the random dropouts in longitudinal studies and the high-dimensionality of big data in modern health research, classification tasks under these complications become even more challenging.

In Chapter 2, we develop a nonparametric approach for the classification of imbalanced data under a longitudinal and high-dimensional structure. The functional principal component analysis is employed to handle the over-time associations within longitudinal profiles and feature extraction is conducted separately for each biomarker. Then, an efficient and easy-to-implement binary classifier is proposed to deal with the imbalanced labels between two classes. Unlike other traditional methods, our approach does not require any distributional assumptions on biomarkers and is capable of performing model estimation and variable selection simultaneously.

In Chapter 3, we further extend the proposed method to multi-class data. For feature extraction, we adopt the natural cubic spline to avoid potential computational burden induced by data density. The proposed approach avoids the construction of multiple binary classifiers and casts the multi-class classification task as a single optimization problem. Besides that, the imbalance issue across different classes has been resolved by the integration of weight-adjusted margins.

In our both methods, the longitudinal/functional data have to be characterized by some techniques in functional data analysis. This step serves as an important rule in our approaches to reduce the longitudinal dimension of data. Two types of techniques have been employed: (1) functional principal component analysis (FPCA) and (2) natural cubic spline (NCS). Technically, FPCA uses pooled data to estimate the covariance and mean function of underlying trajectories. Thus, the prediction of individual smooth trajectory can still be made even if there only exists one or few measurements for a particular subject. However, this estimating process might be time-consuming when the longitudinal data are dense or the total number of subjects is large. In contrast, NCS provides an efficient solution for the feature extraction from functional data. Once the basis functions are determined, a set of spline coefficients can be easily estimated independently for each subject. One limitation with NCS is that it generally requires more longitudinal measurements for spline coefficients estimation. For comparison purpose, we have applied both the FPCA and NCS to the binary Alzheimer's disease data and the numerical results are provided in Table C.1. It seems that the overall performances of FPCA and NCS are close and comparable although NCS achieves a slightly higher AUC on test data. This real data analysis indicates that either FPCA or NCS can be applied for dimension reduction when sufficient data are available for each subject. However, if only few measurements are available, such as one or two, NCS generally fails and FPCA can be used instead.

It should be noted that few studies in the literature take advantage of the underlying associations within longitudinal profiles for the early detection and classification tasks (Tomasko et al., 1999;Marshall and Barón, 2000;De La Cruz-Mesia and Quintana, 2007;Arribas-Gil et al., 2015). Most studies only use data at baseline or from the last visit. Nevertheless, subtle changes may already occur in the over-time progression. Therefore, these features summarised from the longitudinal data using above techniques are expected to be more informative compared to the original features at baseline or from the last visit.

Moreover, the proposed weight-adjusted margin in Chapter 3 can address the issue of imbalance among classes to some extent and lead to an improved classification performance. However, our proposed weights are somewhat simple and naïve. Other advanced weight adjustments can also be considered to improve the classification performance, which is beyond the scope of this study.

It is also worth mentioning that our proposed binary classifier in Chapter 2 is intrinsically different from AdaBoost (Freund and Schapire, 1997) despite both of them adopt the exponential loss function. AdaBoost is essentially a boosting algorithm which constructs a sequence of weak learners and continuously assigns updated weights to misclassified instances. In contrast, we mainly use the univariate exponential loss to approximate the empirical AUC in the objective function, thus leading to improved classification performances under the highly skewed distribution between positive and negative classes.

Lastly, we further investigate the classification performance of our multi-class framework with different numbers of classes. Two versions of gait data are used for this purpose: (1) the original dataset with 42 college-aged subjects (i.e., $C = 42$) and (2) the same dataset with additional 19 subjects (i.e., $C = 61$). As shown in Table B.1, minor improvements have been observed for SVM, LDA, and LMKNCN using the smaller dataset with 42 subjects. On the contrary, our proposed method has obtained a slightly lower overall accuracy with this 42-subject dataset. It still requires additional simulation studies to investigate whether higher numbers of classes will lead to better performances for our approach.

Bibliography

Shivani Agarwal. Surrogate regret bounds for the area under the roc curve via strongly proper losses. In *Conference on Learning Theory*, pages 338–353. PMLR, 2013.

Ana Arribas-Gil, Rolando De la Cruz, Emilie Lebarbier, and Cristian Meza. Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. *Biometrics*, 71(2):333–343, 2015.

Alzheimer's Association. 2021 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 17(3):321–387, 2021.

Richard Baker. Gait analysis methods in rehabilitation. *Journal of neuroengineering and rehabilitation*, 3(1):1–10, 2006.

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.

Milton C Biagioni and James E Galvin. Using biomarkers to improve detection of alzheimer's disease. *Neurodegenerative Disease Management*, 1(2):127–139, 2011.

Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Larry D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, pages 77–82. IEEE, 1994.

Toon Calders and Szymon Jaroszewicz. Efficient auc optimization for classification. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer, 2007.

Shing Chang, Wei-Wen Hsu, and Margaret Rys. An exploratory study to identify key joints for gender differentiation based on walking patterns. *Unpublished Manuscript*, 2020.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

Rolando De La Cruz-Mesia and Fernando A Quintana. A model-based approach to bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hcg profiles. *Biostatistics*, 8(2):228–238, 2007.

Aurore Delaigle and Peter Hall. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286, 2012.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

Marcel Dettling and Peter Bühlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(9):1061–1069, 2003.

Chong-Zhi Di, Ciprian M Crainiceanu, Brian S Caffo, and Naresh M Punjabi. Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1):458, 2009.

José F Díez-Pastor, Juan J Rodríguez, Cesar Garcia-Osorio, and Ludmila I Kuncheva. Random balance: ensembles of variable priors classifiers for imbalanced data. *Knowledge-Based Systems*, 85:96–111, 2015.

Ürün Dogan, Tobias Glasmachers, and Christian Igel. A unified view on multi-class support vector classification. *J. Mach. Learn. Res.*, 17(45):1–32, 2016.

Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605, 2008.

Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

Rong-En Fan, Pai-Hsuen Chen, Chih-Jen Lin, and Thorsten Joachims. Working set selection using second order information for training support vector machines. *Journal of machine learning research*, 6(12), 2005.

Francisco Fernández-Navarro, César Hervás-Martínez, and Pedro Antonio Gutiérrez. A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, 44(8):1821–1833, 2011.

Louis Ferré and Nathalie Villa. Multilayer perceptron with functional inputs: an inverse regression approach. *Scandinavian Journal of Statistics*, 33(4):807–823, 2006.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

Anders M Fjell, Kristine B Walhovd, Christine Fennema-Notestine, Linda K McEvoy, Donald J Hagler, Dominic Holland, James B Brewer, Anders M Dale, Alzheimer's Disease Neuroimaging Initiative, et al. Csf biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and alzheimer's disease. *Journal of Neuroscience*, 30(6): 2088–2101, 2010.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Giovanni B Frisoni, Nick C Fox, Clifford R Jack, Philip Scheltens, and Paul M Thompson. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2): 67–77, 2010.

Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8): 1761–1776, 2011.

Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, 46(12):3460–3471, 2013.

Wei Gao and Zhi-Hua Zhou. On the consistency of auc pairwise optimization. In *IJCAI*, pages 939–945, 2015.

Giovana Gavidia-Bovadilla, Samir Kanaan-Izquierdo, María Mataró-Serrat, Alexandre Perera-Lluna, and Alzheimer's Disease Neuroimaging Initiative. Early prediction of alzheimer's disease using null longitudinal model-based classifiers. *PloS One*, 12(1): e0168011, 2017.

Armineh L Ghazarian, Todd Haim, Samir Sauma, and Pragati Katiyar. National institute on aging seed funding enables alzheimer's disease startups to reach key value inflection points. *Alzheimer's & Dementia*, 2021.

Jeff Goldsmith, Sonja Greven, and CIPRIAN Crainiceanu. Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51, 2013.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

Jianping Gou, Zhang Yi, Lan Du, and Taisong Xiong. A local mean-based k-nearest centroid neighbor classifier. *The Computer Journal*, 55(9):1058–1071, 2012.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

PJ Green and Bernard W Silverman. *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach.* CRC Press, 1993.

TJ Hastie and RJ Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Fabian Horst, Markus Mildner, and Wolfgang I Schöllhorn. One-year persistence of individual gait patterns identified in a follow-up study–a call for individualised diagnose and therapy. *Gait & posture*, 58:476–480, 2017.

Junjie Hu, Haiqin Yang, Michael R Lyu, Irwin King, and Anthony Man-Cho So. Online nonlinear auc maximization for imbalanced data sets. *IEEE transactions on neural networks and learning systems*, 29(4):882–895, 2017.

De-Shuang Huang and Chun-Hou Zheng. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics*, 22(15):1855–1862, 2006.

Meiyan Huang, Wei Yang, Qianjin Feng, and Wufan Chen. Longitudinal measurement and hierarchical classification framework for the prediction of alzheimer's disease. *Scientific Reports*, 7(1):1–13, 2017.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Gareth M James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, 2002.

Gareth M James and Catherine A Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

Gareth M James, Trevor J Hastie, and Catherine A Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, 2000.

Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56. Citeseer, 2000.

Kari Karhunen. *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*, volume 37. Universitat Helsinki, 1947.

Wonjin Kim and Yanggon Kim. Gait recognition for human identification using kinect. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, pages 1–4, 2017.

Ross Kleiman and David Page. Aucμ: A performance metric for multi-class machine learning models. In *International Conference on Machine Learning*, pages 3439–3447. PMLR, 2019.

Stefan Knerr, Léon Personnaz, and Gérard Dreyfus. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing*, pages 41–50. Springer, 1990.

Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In *ICML*, 2011.

UH-G Kreissel. Pairwise classification and support vector machine. *Advances in Kernel Methods*, 1999.

Balaji Krishnapuram, Lawrence Carin, Mário AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005.

Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms.* John Wiley & Sons, 2014.

Jessica BS Langbaum, Kewei Chen, Richard J Caselli, Wendy Lee, Cole Reschke, Daniel Bandy, Gene E Alexander, Christine M Burns, Alfred W Kaszniak, Stephanie A Reeder, et al. Hypometabolism in alzheimer-affected brain regions in cognitively healthy latino individuals carrying the apolipoprotein e $\varepsilon 4$ allele. *Archives of Neurology*, 67(4):462–468, 2010.

Jinny Claire Lee, Soo Jung Kim, Seungpyo Hong, and YoungSoo Kim. Diagnosis of alzheimer's disease utilizing amyloid and tau as fluid biomarkers. *Experimental & Molecular Medicine*, 51(5):1–10, 2019.

Kan Li and Sheng Luo. Dynamic prediction of alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in Medicine*, 38(24): 4804–4818, 2019.

Kan Li, Wenyaw Chan, Rachelle S Doody, Joseph Quinn, and Sheng Luo. Prediction of conversion to alzheimer's disease with longitudinal measures and time-to-event data. *Journal of Alzheimer's Disease*, 58(2):361–371, 2017.

Kan Li, Richard O'Brien, Michael Lutz, Sheng Luo, Alzheimer's Disease Neuroimaging Initiative, et al. A prognostic model of alzheimer's disease relying on multiple longitudinal measures and time-to-event data. *Alzheimer's & Dementia*, 14(5):644–651, 2018.

Yi Li, Juha O Rinne, Lisa Mosconi, Elizabeth Pirraglia, Henry Rusinek, Susan DeSanti, Nina Kemppainen, Kjell Någren, Byeong-Chae Kim, Wai Tsui, et al. Regional analysis of fdg and pib-pet images in normal aging, mild cognitive impairment, and alzheimer's disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 35(12):2169–2181, 2008.

Jeffrey Lin, Kan Li, and Sheng Luo. Functional survival forests for multivariate longitudinal outcomes: Dynamic prediction of alzheimer's disease progression. *Statistical Methods in Medical Research*, page 0962280220941532, 2020.

Yi Lin, Yoonkyung Lee, and Grace Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46(1):191–202, 2002.

Na Liu, Xiaomei Li, Ershi Qi, Man Xu, Ling Li, and Bo Gao. A novel ensemble learning paradigm for medical diagnosis with imbalanced data. *IEEE Access*, 8:171263–171280, 2020.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

Michel Loeve. Functions aleatoires du second ordre. *Processus Stochastique et Mouvement Brownien*, pages 366–420, 1948.

Xiaojing Long, Lifang Chen, Chunxiang Jiang, Lijuan Zhang, and Alzheimer's Disease Neuroimaging Initiative. Prediction and classification of alzheimer disease based on quantification of mri deformation. *PloS One*, 12(3):e0173372, 2017.

Siwei Lyu and Yiming Ying. A univariate bound of area under roc. *arXiv preprint arXiv:1804.05981*, 2018.

Shuangge Ma and Jian Huang. Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21(24):4356–4362, 2005.

Shuangge Ma and Jian Huang. Combining multiple markers for classification using roc. *Biometrics*, 63(3):751–757, 2007.

Qing Mai and Hui Zou. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234, 2013.

Dragos D Margineantu. Class probability estimation and cost-sensitive classification decisions. In *European Conference on Machine Learning*, pages 270–281. Springer, 2002.

Guillermo Marshall and Anna E Barón. Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine*, 19(15):1969–1981, 2000.

Niklas Mattsson, Henrik Zetterberg, Oskar Hansson, Niels Andreasen, Lucilla Parnetti, Michael Jonsson, Sanna-Kaisa Herukka, Wiesje M van der Flier, Marinus A Blankenstein, Michael Ewers, et al. Csf biomarkers and incipient alzheimer disease in patients with mild cognitive impairment. *Jama*, 302(4):385–393, 2009.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Aditya Krishna Menon and Robert C Williamson. Bayes-optimal scorers for bipartite ranking. In *Conference on Learning Theory*, pages 68–106. PMLR, 2014.

Lisa Mosconi. Brain glucose metabolism in the early and specific diagnosis of alzheimer's disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 32(4):486–510, 2005.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys (CSUR)*, 52(2):1–34, 2019.

Ellis Niemantsverdriet, Sara Valckx, Maria Bjerke, and Sebastiaan Engelborghs. Alzheimer's disease csf biomarkers: Clinical indications and rational use. *Acta Neurologica Belgica*, 117(3):591–602, 2017.

Juri Opitz and Sebastian Burst. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*, 2019.

J. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer New York, 2006. ISBN 9780387227511. URL https://books.google.com/books?id=REzuyz_V6OQC.

Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.

Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost:

A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2009.

Hyejin Shin. An extension of fisher's discriminant analysis for stochastic processes. *Journal of Multivariate Analysis*, 99(6):1191–1216, 2008.

Houssem Sifaou, Abla Kammoun, and Mohamed-Slim Alouini. High-dimensional linear discriminant analysis classifier for spiked covariance model. 2020.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1, 2011.

Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

Alex J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

Joan G Staniswalis and J Jack Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418, 1998.

Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.

Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04): 687–719, 2009.

Ke Tang, Rui Wang, and Tianshi Chen. Towards maximizing the area under the roc curve for multi-class classification problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, 2011.

Jafar Tanha, Yousef Abdi, Negin Samadi, Nazila Razzaghi, and Mohammad Asadpour. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7(1):1–47, 2020.

Lisa Tomasko, Ronald W Helms, and Steven M Snapinn. A discriminant analysis extension to mixed models. *Statistics in Medicine*, 18(10):1249–1260, 1999.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Changsheng Wan, Li Wang, and Vir V Phoha. A survey on gait recognition. *ACM Computing Surveys (CSUR)*, 51(5):1–35, 2018.

Xinhuan Wan, Panqin Ma, and Xiangrong Zhang. A promising choice in hypertension treatment: Fixed-dose combinations. *Asian Journal of Pharmaceutical Sciences*, 9(1):1–7, 2014.

Benjamin X Wang and Nathalie Japkowicz. Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, 25(1):1–20, 2010.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

Zhanfeng Wang, Yuan-chin I Chang, Zhiliang Ying, Liang Zhu, and Yaning Yang. A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics*, 23(20):2788–2794, 2007.

Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

Danielle Wilson, Ruth Peters, Karen Ritchie, and Craig W Ritchie. Latest advances on interventions that may prevent, delay or ameliorate dementia. *Therapeutic Advances in Chronic Disease*, 2(3):161–173, 2011.

Lian Yan, Robert H Dodier, Michael Mozer, and Richard H Wolniewicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 848–855, 2003.

Fang Yao, Hans-Georg Müller, Andrew J Clifford, Steven R Dueker, Jennifer Follett, Yumei Lin, Bruce A Buchholz, and John S Vogel. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685, 2003.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

Chun-Wu Yeh, Der-Chiang Li, Liang-Sian Lin, and Tung-I Tsai. A learning approach with under-and over-sampling for imbalanced data sets. In *2016 5Th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 725–729. IEEE, 2016.

Wenbao Yu and Taesung Park. Aucpr: An auc-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC genomics*, 15(10):1–12, 2014.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pages 435–442. IEEE, 2003.

Jun Zhang, Mingxia Liu, Le An, Yaozong Gao, and Dinggang Shen. Landmark-based alzheimer's disease diagnosis using longitudinal structural mr images. In *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*, pages 35–45. Springer, 2016.

Peilin Zhao, Steven CH Hoi, Rong Jin, and Tianbo YANG. Online auc maximization. 2011a.

XG Zhao, W Dai, Y Li, and L Tian. Auc-based biomarker ensemble with an application on gene scores predicting low bone mineral density. *Bioinformatics*, 27(21):3050–3055, 2011b.

XH Zhou, B Chen, YM Xie, F Tian, H Liu, and X Liang. Variable selection using the optimal roc curve: An application to a traditional chinese medicine study on osteoporosis disease. *Statistics in Medicine*, 31(7):628–635, 2012.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77, 2005.

Appendix A

Calculation of gradient and Hessian

A.1 The gradient and Hessian of univariate exponential loss

To compute the hessian of the loss function in Equation (2.4.1) with respect to m, i.e. $\ell^{''}(m)$, it generally requires $O(N^2)$ time complexity, thus leading to intensive computation especially when N is large. According to the argument of Hastie and Tibshirani (1990), $\ell^{''}(m)$ can be replaced by a diagonal matrix with only the diagonal entries of $\ell^{''}(m)$ because the optimal β will remain a fixed point and the off diagonal entries are small compared to the diagonal entries. Let $m = X\beta = [m_1, m_2, ..., m_N]^T$, where $m_i = x_i^T\beta$, and we have the loss function defined as:

$$\ell(\beta) = \sum_{i=1}^{N} e^{-y_i x_i^T \beta}.$$

Thus, it is easy to calculate the gradient as $\ell^{'}(m) = [\ell^{'}(m)_1, ..., \ell^{'}(m)_N]^T$, where $\ell^{'}(m)_i = (-y_i)e^{-y_i m_i}$, and the diagonal entries of the hessian can be computed as

$$\ell^{''}(m)_{i,i} = y_i^2 e^{-y_i m_i}.$$

A.2 The gradient and Hessian of weight-adjusted multi-class exponential loss

To implement the Algorithm 2, partial Newton steps have been employed to conduct partial quadratic approximation to the proposed loss function (3.3.2) defined as follows:

$$\ell_{\Delta}(w) = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{c\neq y_i} e^{\left[-\left(x_i^T w_{y_i} - x_i^T w_c - \Delta_c\right)\right]}\right\}.$$

Let $m = Xw_k = [m_1, ..., m_N]^T$, where $m_i = x_i^T w_k$, it is not difficult to show that entries of the gradient and diagonal entries of the hessian can be computed separately as follows:

$$\ell^{'}(m)_i = \begin{cases} -\sum_{c\neq y_i} \exp(x_i^T w_c - m_i + \Delta_c), & \text{if } y_i = k \\ \exp(m_i - x_i^T w_{y_i} + \Delta_c), & \text{if } y_i \neq k. \end{cases}$$

$$\ell^{''}(m)_{i,i} = \begin{cases} \sum_{c\neq y_i} \exp(x_i^T w_c - m_i + \Delta_c), & \text{if } y_i = k \\ \exp(m_i - x_i^T w_{y_i} + \Delta_c), & \text{if } y_i \neq k. \end{cases}$$

Appendix B

The proposed multi-class framework with different numbers of classes

Table B.1: Results (S.E.) of gait-based person recognition using 42 or 61 subjects (i.e. classes) based on 500 Monte Carlo replicates

Number of subjects			SVM	LDA	LMKNCN	Proposed Method
42	Training	G-Mean	.977(.007)	.999(.001)	.996(.004)	.998(.003)
		Macro Recall	.980(.006)	.999(.001)	.996(.004)	.998(.003)
		Macro Precision	.990(.003)	.999(.001)	.997(.003)	.999(.002)
		Macro F1 Score	.985(.004)	.999(.001)	.997(.003)	.998(.002)
		Overall accuracy	.981(.005)	.999(.001)	.996(.004)	.998(.002)
	Test	Overall accuracy†	.764(.054)	.936(.035)	.814(.054)	.869(.050)
61	Training	G-Mean	.956(.010)	.999(.001)	.989(.004)	.999(.001)
		Macro Recall	.964(.006)	.999(.001)	.990(.003)	.999(.001)
		Macro Precision	.980(.004)	.999(.001)	.992(.003)	.999(.001)
		Macro F1 Score	.972(.005)	.999(.001)	.991(.003)	.999(.001)
		Overall accuracy	.968(.005)	.999(.001)	.990(.003)	.999(.001)
	Test	Overall accuracy†	.727(.048)	.913(.033)	.788(.047)	.907(.036)

LMKNCN: local mean-based k-nearest centroid neighbor; †: calculated based on only one observation.

Appendix C

Comparison between FPCA and NCS using binary Alzheimer's disease data

Table C.1: Classification results (S.E.) of our proposed binary classifier on ADNI data over 500 Monte Carlo replicates using either FPCA or NCS for feature extraction

		FPCA	NCS
Training Set	Sensitivity	.946(.066)	.908(.048)
(n_h=166, n_d=21)	Specificity	.973(.035)	.909(.040)
	Accuracy	.970(.035)	.909(.032)
	AUC	.976(.033)	.960(.013)
Test Set	Sensitivity	.790(.145)	.819(.127)
(n_h=71, n_d=9)	Specificity	.880(.094)	.893(.047)
	Accuracy	.870(.084)	.885(.040)
	AUC	.880(.091)	.945(.036)

FPCA: functional principal component analysis; NCS: natural cubic spline