ORIGINAL ARTICLE
Journal Section

Review on Functional Data Classification

Shuoyang Wang[1*] | Yuan Huang[1*] | Guanqun Cao[2†]

[1]Department of Biostatistics, Yale University, U.S.A.

[2]Department of Statistics and Probability, Michigan State University, U.S.A.

Correspondence
Guanqun Cao, Department of Statistics and Probability, Michigan University, U.S.A.
Email: caoguanq@msu.edu

Funding information
Cao's research was also partially supported by Simons Foundation under Grant #849413 and National Science Foundation under Grants CNS-2319342 and CNS-2319343.

A fundamental problem in functional data analysis is to classify a functional observation based on training data. The application of functional data classification has gained immense popularity and utility across a wide array of disciplines, encompassing biology, engineering, environmental science, medical science, neurology, social science, and beyond. The phenomenal growth of the application of functional data classification indicates the urgent need for a systematic approach to develop efficient classification methods and scalable algorithmic implementations. Therefore, we here conduct a comprehensive review of classification methods for functional data. The review aims to bridge the gap between the functional data analysis community and the machine learning community, and to intrigue new principles for functional data classification.

1 | INTRODUCTION

Functional data refer to curves or functions, wherein the data corresponding to each variable are represented as smooth curves, surfaces, or hypersurfaces, evaluated at a finite subset of some intervals. These intervals may exist in one dimension (1D), two dimensions (2D), or three dimensions (3D), such as a time period, a range of pixels or voxels, and other applicable contexts. Frequently encountered forms of functional data encompass trajectories, time series data, spatio-temporal data, and various other representations. Despite the diverse forms functional data may take, their key characteristic lies in the existence of underlying functions, such as curves, which define the essential nature of the data.

Compared with classical independent and identically distributed (i.i.d.) data, the distinguishing feature of functional data is the presence of dependence and smoothness within each data curve. Instead of dealing with random variables, functional data involves random processes, which necessitates the implicit consideration of an infinite-dimensional function space. Functional data analysis (FDA) has become a topic of growing interest within the statistics community over recent decades, leading to a wealth of literature on the subject. For a comprehensive understanding of FDA theory and methods, we refer the readers to the classic textbook by Ramsay and Silverman [2005], recent monographs [Ferraty and Romain, 2011, Hsing and Eubank, 2015] and review papers Wang et al. [2016a], Morris [2015]. Recently, to extend classification approaches from multivariate data to functional ones, Jiang and Chen [2020] reviewed the fundamental theory and methods on how to leverage dimension reduction (filtering) methods for functional data settings. In this paper, we present several emerging and promising areas of functional data classification that we anticipate will gain momentum in the near future and have a progressively significant impact across various scientific domains.

The rest of this paper proceeds as follows. In Section 2, several motivation examples, including 1D, 2D and 3D functional data, are provided. In Section 3, we introduce the preliminary ideas and notations for functional data and classification. Reviews of density-based, regression-based, distance based, SVMs, and reproducing kernel methods are provided in Section 4. Section 4 also includes review of classification methods for multivariate functional data and robust classifiers. Evaluation criteria for functional data classifiers have been reviewed in Section 5. We include a table of several open-source software and packages on functional data classification in Section 6. The paper concludes with summary and discussions as provided in Section 7.

2 | MOTIVATION EXAMPLES

2.1 | 1D functional data

2.1.1 | TIMIT speech data

The TIMIT speech data was derived from the TIMIT Acoustic-Phonetic Continuous Speech Corpus database, which was compiled by the National Technical Information Service under the United States Department of Commerce. This database has garnered widespread recognition and serves as a vital resource for advancing research in speech recognition and functional data classification [Ferraty and Vieu, 2003, Delaigle and Hall, 2012, Wang et al., 2023b]. From the digitized speech contained in the TIMIT database, five phonemes were extracted and transcribed as follows: "sh" as in "she", "dcl" as in "dark", "iy" as the vowel sound in "she", "aa" as the vowel sound in "dark", and "ao" as the initial vowel sound in "water". Prior to analysis, a log-periodogram transformation was applied to each speech frame. This transformation facilitates the representation of speech data in a suitable format for speech recognition. Moreover, each speech frame is represented by 400 samples, obtained through a 16-kHz sampling rate. From these samples, the first 150 frequencies were retained for further analysis. To provide visual insight, Figure 1 presents 30 log-periodograms for four phoneme classes, allowing for a comparative display of acoustic properties across the selected phonemes.

2.1.2 | DTI data

The DTI dataset, available within the *refund* package in R, provides researchers with a valuable resource of diffusion tensor imaging (DTI) data. DTI is an advanced imaging technique widely employed in neuroscience and medical research to investigate the intricate structure and connectivity of white matter in the brain. Of particular interest in the analysis of this dataset is classification of individuals into two distinct groups: those diagnosed with multiple sclerosis (MS) and a control group. The classification is based on the examination of fractional anisotropy tract profiles specifically pertaining to the corpus callosum (cca) and the right corticospinal tract (rcst). By analyzing the tract profiles

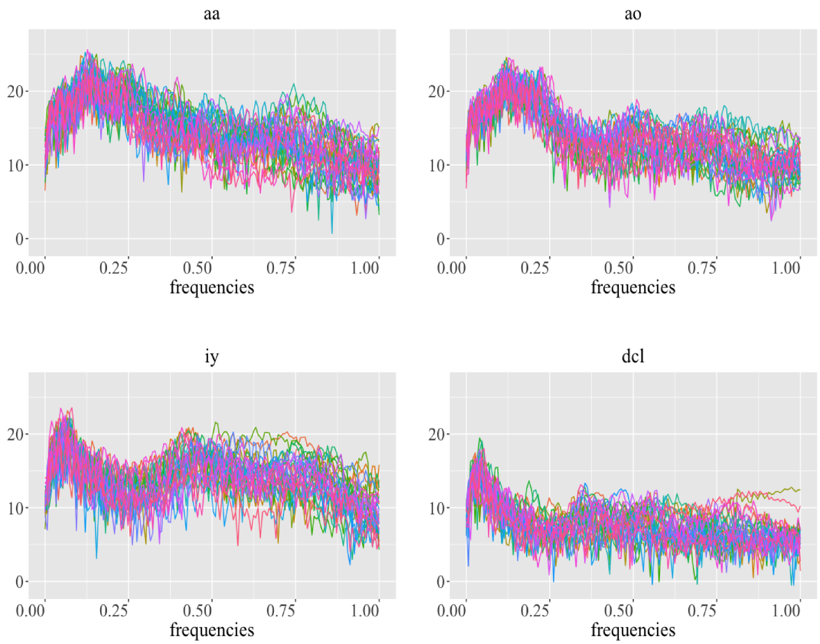

FIGURE 1 Illustration of TIMIT speech data: a sample of 30 log-periodograms for each of the "aa", "ao", "iy", and "dcl" phonemes.

within these specific regions, researchers can gain insights into the distinct characteristics and abnormalities present in the white matter pathways associated with MS. This classification task bears great importance in comprehending the implications of MS on the integrity and connectivity of these neural pathways. See Figure 2 for data illustration of a randomly selected set of 30 individuals.

2.2 | 2D/3D functional data

2.2.1 | MNIST handwritten digits

This 2D functional data example was extracted from the MNIST database (`http://yann.lecun.com/exdb/mnist/`). This classical MNIST database contains 60,000 training images and 10,000 testing images of handwritten digits $(0, 1, \dots, 9)$. These images were normalized and centered to fit into a 28×28 pixel bounding box, and anti-aliased. As each individual pixel assumes a single value that signifies the intensity level, spanning the range from 0 to 255, it is natural to conceptualize the image as a function endowed upon a square domain. This inherent functional nature allows for the direct construction of functional data classifiers based on samples. Figure 3 portrays a representative sample of the MNIST dataset, emphasizing the varying intensity levels. Wang et al. [2023a] developed a deep neural network based classifier to conduct multi-class classification for this dataset.

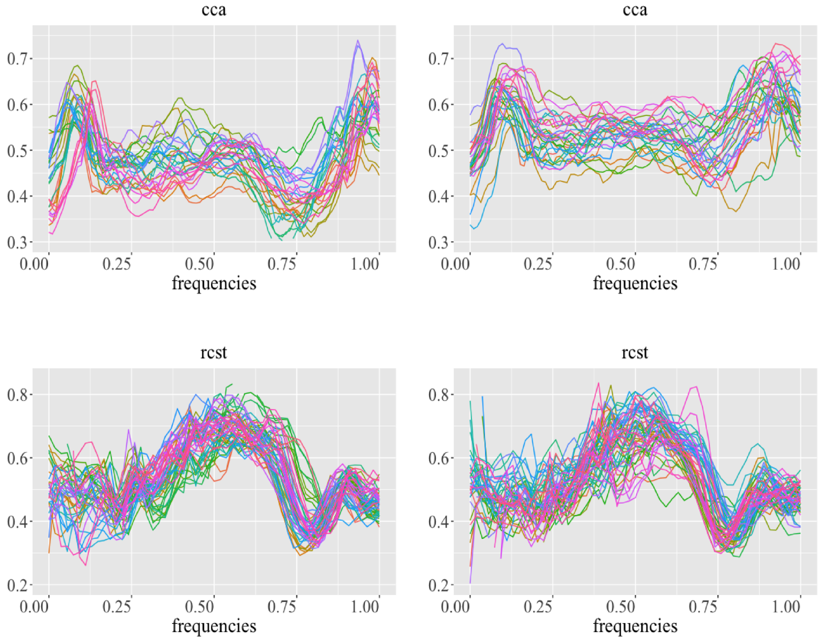

FIGURE 2 Illustration of DTI data: a sample of 30 fractional anisotropy tract profiles for corpus callosum (cca) and right corticospinal tract (rcst). Left: MS cases. Right: healthy control.

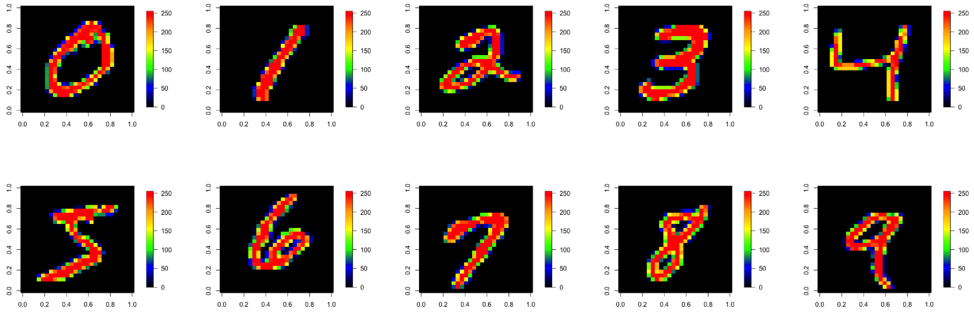

FIGURE 3 Illustration of MNIST data.

2.2.2 | ANDI PET imaging data

Another illustrative instance of multidimensional functional data can be found in the positron emission tomography scans extracted from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database available at `adni.loni.usc.edu`. The ADNI database, which encompasses a longitudinal multicenter study, is specifically designed to advance the development of clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking

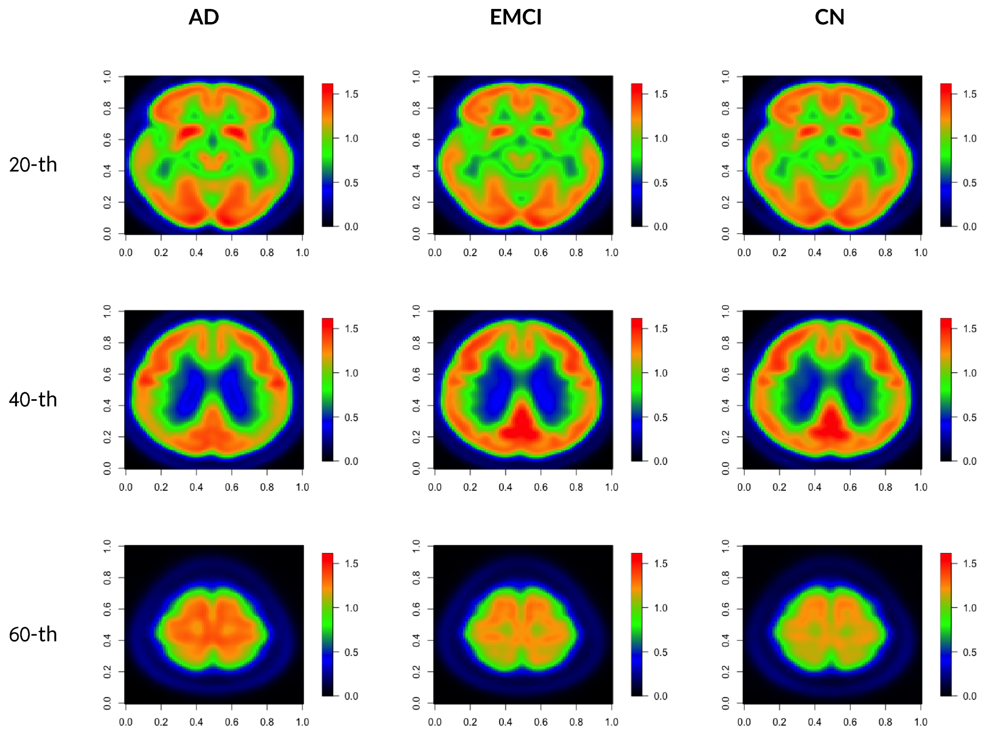

FIGURE 4 Illustration of ANDI data. Averaged images of the 20-th, the 40-th and the 60-th slices of AD group (left), EMCI group (middle), and CN group (right).

of Alzheimer's disease (AD). A compelling approach to differentiate the three groups representing different stages of AD progression, namely the control group (CN), individuals with early mild cognitive impairment (EMCI), and diagnosed AD patients, involves the application of binary/multiclass classification techniques to the PET imaging data. In Figure 4, we provide a visual representation of the averaged imaging data across various groups and slices. Each image has undergone spatial normalization and post-processing procedures. All scans have been standardized to a $79 \times 95 \times 68$ voxel format such that each patient possesses 68 2D image slices, each containing 79×95 pixels. For 2D analysis, each selected image slice contains $79 \times 95 = 7,505$ observed pixels. In case of 3D analysis, the number of observed voxels in each patient's brain sample amounts to $79 \times 95 \times 68 = 510,340$. Wang et al. [2023a] have proposed multi-dimensional functional data classification methods for both 2D and 3D ADNI data.

3 | PRELIMINARIES

Conventionally, functional data is a collection of independently and randomly observed curves which are real-valued functions [Wang et al., 2016b]. Let $X(t)$ be a random process residing in $\mathcal{X}$ with mean function $\mu(t) = \mathrm{E}[X(t)]$ and covariance function $\Omega(t, t') = \mathrm{Cov}(X(t), X(t'))$ for $t, t' \in \mathcal{T}$, which is typically considered in a Hilbert space, such as $L_2(\mathcal{T})$ and $\mathcal{T} \subset \mathbb{R}$. Generally speaking, the next-generation functional data can be extensively denoted as $X(\boldsymbol{t})$ for some $\boldsymbol{t} \in \mathcal{T} \subset \mathbb{R}^d$, $d \geq 2$. This generalization finds widespread application in various fields, including imaging data, where the intensity value linked to each pixel is considered the value of a function at the corresponding spatial location such that each image can be viewed as a realization of a random function. However, there is lack of literature on such next-generation functional data classification. Hence, in the following, we mainly focus on the one-dimensional ($d = 1$) functional data classification, unless we specify the value of dimension d.

3.1 | Formulation of functional data classification

In statistics, classification involves determining which specific group a new observation belongs to, using a training data set that includes samples with labeled groups. As a form of supervised learning, classification garners significant attention in both statistics and machine learning communities due to its wide-ranging applications. For integer $K \geq 2$, we consider a K-class classification problem with the functional observations defined on space $\mathcal{X}$. Let (X_i, Y_i), $i = 1, \ldots, n$, be i.i.d. random pairs of observations, where $X_i \in \mathcal{X}$ and $Y_i \in \{1, \ldots, K\}$. For a new observation X, the classification task is to predict the class label Y by a classifier $C : \mathcal{X} \to \{1, \ldots, K\}$, based on finite sample $\{(X_i, Y_i)\}_{i=1}^n$. More specifically, the classification rule based on finite sample is defined as $\widehat{C}_n \equiv \widehat{C}_n\left((X_1, Y_1), \ldots, (X_n, Y_n)\right)$. We denote the prior probability $\pi_k = \mathbb{P}(Y = k)$, and the posterior probabilities $p_k(x) = \mathbb{P}(Y = k|X = x)$, $k = 1, \ldots, K$.

3.2 | Dimension reduction

Compared to the classical multivariate data, the inherent infinite-dimensionality of functional data necessitates the utilization of dimension reduction techniques to effectively process functional observations. These techniques are imperative for almost all functional statistical tasks, including functional data classification. There are generally two main strategies for processing functional data in classification tasks. The first involves dimension reduction techniques on functional observation, which transform the functional data into a lower-dimensional representation that can be analyzed using traditional multivariate data classification methods. Examples of this category include Functional Principal Component Analysis (FPCA) [Müller, 2005, Leng and Müller, 2006] , basis expansion [Wang et al., 2023a,b], partial least squares Preda et al. [2007], and other similar techniques. The second category involves methods that preserve the full continuum of the functional data. These methods operate directly on the functional data without reducing it to a lower-dimensional representation first. However, depending on the specific method employed, some may still require dimension reduction. Examples of this category include distance-based methods and methods that leverage the special function space structure, such as Reproducing Kernel Hilbert Space (RKHS) [Berrendero et al., 2018, Sang et al., 2022].

As the most widely recognized and prevalent dimension reduction method for functional data, FPCA is highly regarded for its exceptional effectiveness and interpretability. Let $X(t)$, $t \in \mathcal{T} := [0, 1]$ be a random process with $\int_{\mathcal{T}} \mathbb{E} X(t)^2 dt < \infty$. $X(t)$ has an unknown mean function $\mu_k(t)$ and unknown covariance function $\Omega(t, t')$, for $t, t' \in \mathcal{T}$. By Mercer's theorem [Mercer, 1909], the covariance function $\Omega(t, t')$ can be represented through the spectral decomposition

$$\Omega(t, t') = \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \phi_j(t'),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are nonnegative eigenvalues satisfying, $\sum_{j=1}^{\infty} \lambda_j < \infty$, and $\phi_k(t)$ is the corresponding orthonormal eigenfunction, i.e., $\int_{t \in \mathcal{T}} \phi_j(t) \phi_k(t) dt = \mathbb{I}(j = k), j, k, \in \mathbb{N}^+$. Then, given n random curves $\{X_1(t), \ldots, X_n(t)\}$, by

Karhunen-Loève expansion, $X_i(t)$ can be rewritten as

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_{ij}\phi_j(t), \quad i = 1, \ldots, n,$$

where $\xi_{ij} = \int_{\mathcal{T}} (X_i(t) - \mu(t))\phi_j(t)dt$ are uncorrelated variables with mean 0 and variance λ_j, and ξ_{ij} are called Functional Principal Component Scores (FPCS). In practice, FPCA leverages a truncated expansion using either a prefixed function basis, such as wavelet basis [Zhu et al., 2012], Fourier basis [Wang et al., 2023b], spline basis [James and Hastie, 2001], or embraces a data-adaptive eigenbasis [Delaigle and Hall, 2012, Dai et al., 2017]. This versatility in choosing the basis allows FPCA to offer more flexibility and customization in representing functional data.

4 | METHODOLOGIES FOR FUNCTIONAL DATA CLASSIFICATION

In this section, we organize the functional classification into five main categories. In Section 4.1, we review the density-based classifiers, where both Gaussian and non-Gaussian types of data are discussed. In Sections 4.2 and 4.3, we review the regression model-based and distance-based classifiers, including two major methods: centroid and nearest neighbor, followed by discussions of the functional linear regression approach on the use of different link functions. The RKHS will be discussed in Section 4.4. In Section 4.5, we further discuss the support vector machine which can detect the classification boundary. Finally, we review several recently developed methods tailored for multivariate and multidimensional functional data. These methods exemplify the continuous advancements in the field and hold the potential to shape the future of functional data classification.

4.1 | Density-based classification

First proposed by Fisher [1936], discriminant analysis is a fundamental statistical technique to classify observations into pre-defined groups based on their measured characteristics or variables. This approach is motivated by the principle of maximum likelihood, which seeks the parameter values that make the observed data most probable under the assumed distribution. By assuming that the underlying data distribution follows a specific probabilistic form, such as Gaussianity, the discriminant function is pre-specified and the goal of the training process is to estimate the parameters of this function. When confronted with violations of Gaussianity, parametric density estimation techniques may not be adequate, and nonparametric methods such as kernel density estimation and neural networks can provide more flexible alternatives. By utilizing nonparametric approaches, the resulting density estimates are not restricted to any particular functional form, making them more adaptable to complex and diverse data distributions. In the rest of this subsection, we review the development of functional classification methods under both Gaussian and non-Gaussian assumptions.

4.1.1 | Functional observations with the Gaussian assumption

Under the Gaussian assumption, both observations in the time domain and the transformed correspondence in the frequency domain are Gaussian. One common example is by assuming the first J FPCS $\boldsymbol{\xi}_J = (\xi_1, \ldots, \xi_J)^\top$ satisfies

$$\boldsymbol{\xi}_J | Y = k \sim N(\boldsymbol{\mu}_k, \Lambda_k),$$

and the classifier is defined as

$$\underset{k=1\ldots,K}{\arg\min}\,(\boldsymbol{\xi}_J - \boldsymbol{\mu}_k)\,\Lambda_k^{-1}\,(\boldsymbol{\xi}_J - \boldsymbol{\mu}_k) + \log(|\Lambda_k|) - 2\log\pi_k.$$

James and Hastie [2001] first considered the representation of functional observations using splines basis, and they applied multivariate Linear Discriminant Analysis (LDA) to the spline coefficients to construct the classifier. In order to overcome the limitations imposed by the spline basis, Delaigle and Hall [2012] advocated the centroid method through the incorporation of FPCA. As a data-driven technique, it is suited to capture the underlying patterns of the data. This method is essentially equivalent to Functional LDA (FLDA) [James and Hastie, 2001] under certain scenarios. Subsequently, Delaigle and Hall [2013] proposed the centroid method in functional Quadratic Discriminant Analysis (FQDA) to further address heteroscedasticity. Similar FQDA procedure can be found in Wang et al. [2023b], where they employed the Fourier basis to extract the FPCS. Yu et al. [2022] focused on non-stationary Gaussian process, where they vectorized the functional observations as high-dimensional Gaussian data in a multivariate setup and performed posterior inference to establish the classifier. Gaussianity is crucial for the conventional LDA and quadratic discriminant analysis (QDA), and same conclusion can be obtained in functional classification. When Gaussian assumption is invalid, these functional LDA and QDA based classifiers incur large empirical risks. For example, Wang et al. [2023b] evaluated the numerical performance of their proposed classifier with popular functional QDA under non-Gaussian process situations. The functional QDA has larger empirical misclassification error rates than the proposed classifier which is free of the assumption of the Gaussianity.

4.1.2 | Distribution-free discrimination analysis

In the absence of the Gaussian assumption, nonparametric methodologies are often adopted. One of the most commonly used scheme is to approximate $p_k(x)$ by its finite counterpart

$$p_k \approx \frac{\pi_k f_k(\boldsymbol{\xi}_J)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{\xi}_J)},$$

where J is the number of truncation, $\boldsymbol{\xi}_J$ are the first J selected FPCS, f_k is the density function given X in the k-th class, $k = 1, \ldots, K$. Hall et al. [2001], Bongiorno and Goia [2016], Dai et al. [2017], and Zhang and Sakhanenko [2019] employed the kernel density estimation methods to obtain $\left\{\widehat{f}_k(\boldsymbol{\xi}_J)\right\}_{k=1}^{K}$ by considering independent FPCS, where $\widehat{f}_k(\boldsymbol{\xi}_J) = \prod_{j=1}^{J} \widehat{f}_{kj}(\xi_j)$, and $\widehat{f}_{kj}$ are density function estimators for the j-th FPCS in the k-th group. To overcome the restrictive independence assumption, Wang et al. [2023a] additionally took into account a broader scenario encompassing situations where intricate $\{p_k\}_{k=1}^{K}$ are present. Particularly, p_k incorporate convoluted interactions among FPCS. To tackle the complexity, they utilized the powerful deep neural networks (DNN) to construct the classifiers. Interestingly, due to the computational burden, most of the nonparametric discriminant analysis works have to assume independence among the features. However, it is natural to assume the independence between principal components for functional data classification, which can be seen as a distinctive advantage of FDA.

To account for the inherent continuity of functional data, Ferraty and Vieu [2003] proposed a kernel-type estimator $\widehat{p}_{k,h}(x) = \frac{\sum_{i=1}^{n} \mathbb{I}(Y_i=k)\Lambda\left(h^{-1}d(X_i,x)\right)}{\sum_{i=1}^{n} \Lambda(h^{-1}d(X_i,x))}$, where Λ is the kernel function, h is the bandwidth, and $d(X_i, x)$ is some semi-metric to measure distance between functional observation X_i and x. Unlike the direct estimation of $\{f_\ell(\boldsymbol{x}_J)\}_{\ell=1}^{K}$, they considered $L^2(\mathcal{T})$ space, where X admits the FPCA expansion, and the distance $d(X_i, x)$ is thus approximated by its truncated version $d_J(X_i, x) = \sqrt{\sum_{j=1}^{J}\left[\int_{\mathcal{T}} (X_i(t) - x(t))\,\phi_j(t)\,dt\right]^2}$. Based on the same idea, Chang et al. [2014]

considered a wavelet-based distance with certain thresholds, and the proposed classifier has better performance for noisy image data in their numerical studies. In Meister [2016], the functional data are considered as realisations of random variables which take their values in a general Polish metric space. They further imposed certain metric entropy constraints on this space; but no algebraic properties are required.

4.2 | Regression-based methods

A regression model is constructed linking class identities with functional predictors, frequently through generalized linear models. The model parameters are estimated and used for classification. In general, functional linear regression constructs the model on the inner product $\langle \beta, X \rangle$, where $\beta \equiv \beta(t)$ is the decision boundary to be estimated. It also inherently intertwines with the discriminant analysis, where β is considered as the discrimination curve. This scheme enables direct applicability of multivariate classification techniques, making it a powerful tool for analyzing the infinite-dimensional functional data. Commonly, the dimension reduction is employed on the inner product, where the $\int_{\mathcal{T}} \beta(t) X(t) dt$ is estimated by its finite counterpart.

Generalized functional linear regression is an extension of functional linear regression that allows for the modeling of functional data with labeled responses. For instance, one can model functional logistic regression as

$$\log \left(\frac{p_1(X)}{p_2(X)} \right) = \alpha + \int_{\mathcal{T}} \beta(t) X(t) dt,$$

where $\alpha \in \mathbb{R}$ and $X(\cdot), \beta(\cdot) \in L_2$. Müller [2005] considered the generalized functional linear model for the sparse functional data, where the integral is estimated by the first J truncated FPCS. With the similar idea, Leng and Müller [2006] applied the functional logistic regression to temporal gene expression data. Preda et al. [2007] used partial least square regression to estimate the discriminant coefficient functions. Araki et al. [2009] considered a finite representation using Gaussian basis functions, where the regularization has been utilized to control the smoothness and sparsity. Li and Ghosal [2018] considered unordered multinomial probit, ordered multinomial probit and multinomial logistic regression models via Bayesian approach, where the conditional probability p_k is approximated by some cumulative distribution function of $\int_{\mathcal{T}} \beta(t) X(t) dt$. Müller [2005] considered the so-called generalized functional linear model to the case of sparse longitudinal predictors through the FPCA method. Goldsmith et al. [2011] extended Müller [2005]'s work by proposing a penalized functional regression approach. They smoothed the covariance operators using a large number of eigenvectors to capture the variability of the functional predictors, and modeled the functional regression parameters as penalized splines. This provided a unified framework for functional regression in many settings, including when functions are measured with errors, at equal or unequal intervals, at a dense or sparse set of points, and to multiple functional regressors observed at one or multiple levels. Zhu et al. [2010] performed classification of complex, high-dimensional functional data using the Functional Mixed Model (FMM) framework. The FMM relates a functional response to a set of predictors through functional fixed and random effects, which allows it to account for various factors and between-function correlations.

4.3 | Distance-based approaches

Distance-based classifiers are a popular category of classification methods that rely on measuring the distances between observations in a feature space to make predictions. They become particularly useful in functional classification problems, where the distances between observations are typically unique to the associated function space. Distance-based classifiers find significant utility in functional classification problems because functional data often has a potentially infinite-dimensional structure and measuring the distances between observations can be an effective way to capture the underlying variability and structure of the data. As will be discussed further below, distance-based methods, including centroid classifiers and nearest neighbor, serve as valuable tools for analyzing functional data.

4.3.1 | Centroid classifier

The centroid classifier calculates the distance between the functional observations and the centroids of each classes, which is conceptually simple and interpretable. Specifically, the centroid for the k-th class is given by $\bar{X}_k = n_k^{-1}\sum_{i=1}^{n} X_i \mathbb{I}(Y_i = k)$, and the centroid classifier assigns the new data X to the group with the minimal $d(X, \bar{X}_k)$. Delaigle and Hall [2012] and Delaigle and Hall [2013] proposed the functional centroid classifier, where the distance is constructed by the absolute difference of inner products. See Galeano et al. [2015] for a similar setup in conjunction with the functional Mahalanobis semidistance. Li and Xiao [2019] utilized FLDA to refine their design process by identifying the optimal sampling time points for accurately classifying functional data. Kraus and Stefanucci [2019] reformulated the centroid method as an optimization problem and searched the solution by the conjugate gradient method with early stopping. Chen and Jiang [2018] proposed the sensible FLDA to address the situation when the functional data is generated by a model with multiplicative random effects, such that both the between- and within-class covariance functions were taken into consideration. Darabi and Hosseini-Nasab [2020] introduced a weighted version of the centroid classifier that is based on projection functions and can lead to asymptotically perfect classification.

4.3.2 | Nearest neighbor classifier

The nearest neighbor classifier, another notable distance-based classifier for multivariate data, is frequently employed in classification tasks wherein the objective is to prognosticate the class label of an incoming observation by drawing on the labels of its neighboring data points. The concept of the nearest neighbor algorithm is inherently transferrable to functional data as the distance metric $d(\cdot,\cdot)$ used by the classifier can be expressed in terms of the distance between functions. Specifically, one chooses q closest data $X_{(1)}, \dots, X_{(q)}$ by computing the distances $d(X, X_i)$, and the class label of X is decided by the majority vote of the labels of q instances. The definition of the k-NN classifier can be easily translated to the functional setup by replacing the usual Euclidean distance in $\mathbb{R}^d$ with an appropriate functional metric D. For example, Galeano et al. [2015] defined the functional Mahalanobis semi-distance between $X_1(t)$ and $X_2(t)$ as $D(X_1, X_2) = \left(\sum_{k=1}^{K}(\omega_{1k} - \omega_{2k})^2\right)^{1/2}$, where ω_{ik}, $i = 1, 2$ and $k = 1, 2, \dots, K$, are the standardized FPCS of $X_1(t)$ and $X_2(t)$, respectively. Another straightforward implementation strategy is to perform nearest neighbor classification based on the finite number of projection coefficients or FPCS directly. The inherent robustness of the nearest neighbor classifier against extreme data points renders it a highly valuable alternative to the centroid classifier. However, this advantage comes at the cost of an increased computational burden. Notably, when considering the binary classification, the nearest neighbor procedure can be seen as a case of the plug-in estimator of $p_k(x)$, such that

$$\widehat{p}_k(x) = q^{-1}\sum_{i=1}^{n} \mathbb{I}\left(X_i \in \delta_q(x)\right) Y_i,$$

where $\delta_q(x)$ is the set of q nearest neighbors of x.

Biau et al. [2005] used the first coefficients of a Fourier series expansion of $X_i(t)$, and perform nearest neighbor classifier under a finite dimension. Cérou and Guyader [2006] explored the nearest neighbor classifier in infinite dimensional metric space, and argued that the weak convergence of nearest neighbor classifier holds only if the metric space is separable with some regular conditions. Subsequently, Biau et al. [2010] considered some separable Banach space and provided various examples such as Sobolev spaces, Besov spaces, and RKHS. Baíllo and Cuesta-Albertos [2011] further showed the consistency of the nearest neighbor rule for the Gaussian families. Galeano et al. [2015] established the nearest neighbor classifier using functional Mahalanobis semidistance.

4.4 | Reproducing kernel

RKHS is a special type of Hilbert space that is equipped with a reproducing kernel, which is a positive definite function that plays a role similar to the inner product in the Hilbert space. Particularly, for a Hilbert space $\mathcal{H}$ of functions defined on domain $\mathcal{T}$, a reproducing kernel $r(\cdot,\cdot)$ satisfies $X(t) = \langle r(t,), X\rangle$ for $\forall X \in \mathcal{H}$. Equipped with the reproducing kernel, one can define $\mathcal{H}$ as a RKHS if the span of $r(\cdot,\cdot)$ is dense in $\mathcal{H}$. The reproducing kernel allows for the efficient computation of inner products between functions, which is important for many machine learning algorithms that involve computing distances or similarities between data. The RKHS allows for a more flexible and powerful representation of functions, and the use of kernel-based methods offers a more efficient way to deal with the intrinsically infinite-dimensional functional data.

Shin [2008] extended the multivariate Fisher's linear discriminant analysis, which is based on the spectral decomposition of the linear operator defined on RKHS associated with a second-order stochastic process. Yao et al. [2016] used the representer theorem to construct the decision function with a finite sample. Berrendero et al. [2018] provided a formalized theory for binary functional classification, and the system was established by explicitly calculating the Radon-Nikodym derivatives between Gaussian measures that are absolutely continuous. Torrecilla et al. [2020] further extended the homoscedastic setting to a heteroscedastic setting of Gaussian processes, and discussed the optimal discrimination rules more comprehensively. Berrendero et al. [2023] explored the RKHS as an alternative formulation to the L_2-based model for functional logistic regression. Sang et al. [2022] extended the multivariate distance-weighted discrimination based classifier under RKHS.

4.5 | Support vector machine

As an important algorithm in machine learning, the extension to functional Support Vector Machine (SVM) aims to find the decision boundary $\beta(\cdot)$ that best separates the different classes of functional observations by maximizing the margin. Rossi and Villab [2006] first introduced SVM to classify functional data with linear decision boundaries. In addition, the functional-adjusted kernels was also taken into consideration to address the nonlinear separation. Li and Yu [2008] applied the SVM to a LDA-projected subspace, where the reduced-dimensional subspace is obtained by conventional LDA. Wu and Liu [2013] used the robust truncated-hinge-loss SVM by truncated FPCS to handle sparse and irregularly observed functional data and longitudinal data. The approach proposed by Yao et al. [2016] involved extending probability-enhanced effective dimension reduction to functional predictors by using functional cumulative slicing. Subsequently, they applied the resulting surrogates to a weighted SVM.

4.6 | Robust functional data classification methods

The development of robust classifiers is motivated by the pursuit of acquiring reliable estimates for class memberships. Cuevas et al. [2006] first introduced the concept of data depth into functional data classification. The idea of data depth, which has been investigated in i.i.d. data and multivariate data, has a notable advantage in developing robust estimators for a “location parameter” in high-dimensional or functional frameworks. For instance, when the definition of depth is available, definition of robust estimators such as the median and trimmed means is straightforward. In terms of classification, the utilization of data depth is to assign a coming datum according to its relative depth in the training samples. Cuevas et al. [2006] considered five different proposals on data depth, namely, the Fraiman-Muniz Depth [Fraiman and Muniz, 2001], the h-mode depth, and three random projection methods inspired by the work of Cuesta-Albertos et al. [2007]. Some other depthed-based classification methods are proposed in López-Pintado and Romo [2005], Kwon et al. [2016], Sguera et al. [2014], Hlubinka et al. [2015], and Hubert et al. [2015]. Beyaztas and Shang [2022] proposed a robust functional partial least squares (FPLS) method which produces robust estimates of the regression coefficients in a scalar-on multiple-function regression model. The M-estimator and Tukey’s bisquare loss function were used to solve the regression problem of a scalar response on the extracted FPLS components and to approximate the regression coefficient function. Zhu et al. [2012] proposed a robust, wavelet-based functional mixed model which permits potentially heavier tails for features of the functions indexed by particular wavelet coefficients, leading to a down-weighting of outliers that consequently makes the method robust to outlying functions or regions of functions.

4.7 | Classifiers for multivariate functional data

A multivariate functional datum is defined as a finite dimensional vector where each component is a univariate function. A simple approach to handle functional data consists of discretization of the multivariate function then apply multivariate techniques to the resulting vector. However, functional data are intrinsically infinite dimensional and have properties that differentiate them from multivariate data. Examples of multivariate functional cases include gait data and hand writing data [Ramsay and Silverman, 2005], height and weight of children by age [López-Pintado et al., 2014] and electrocardiogram data [Dai and Genton, 2019].

The variation between different groups of curves in functional data classification usually arises from variations in the data’s diverse patterns or shapess. By extending the depth-based scalar outlyingness to an outlyingness matrix, which contains pure information of shape variation of a curve, Dai and Genton [2018] proposed classifiers for both univariate and multivariate functional data. Moindjié et al. [2022] considered partial least square classification and tree partial least square-based methods for multivariate functional data which are defined in different domains. Blanquero et al. [2019] considered the time as a continuous variable, and they searched for the global solution using a surrogate of the number of misclassified data, namely the correlation between the SVM score and the actual class. When predictors consist of multiple functions, some of which may be redundant. We also note that in functional data classification, functional observations are often contaminated by various systematic effects. These effects may lead to classification bias. Hence, selecting a subset of the functions can reduce the cost of data collection for future observations, and may improve classification accuracy. Zhu et al. [2010] proposed a Bayesian hierarchical model with selection of functional predictors for complex functional data classification problems, where multiple functional predictors are influenced by random batch effects and fixed effects.

5 | EVALUATION OF FUNCTIONAL CLASSIFIERS

When evaluating a classifier, it is customary to employ two principal benchmarks. The first benchmark entails evaluating the misclassification risk

$$R(\widehat{C}_n) = \mathbb{P}\left(Y \neq C(X) | (X_1, Y_1), \ldots, (X_n, Y_n)\right),$$

which utilizes a straightforward and intuitive method that quantifies the probability of erroneous classifications. The second benchmark involves analyzing the consistency of the classifier with the Bayes classifier, a highly regarded approach that provides a measure of the classifier's proximity to the optimal classification strategy, thereby offering greater reliability and insight into the performance. Particularly, a classifier is consistent if its misclassification risk converges to the Bayesian risk, such that

$$\mathrm{E}\, R(\widehat{C}_n) - R(C^*) \to 0, \text{as } n \to \infty,$$

where $C^*(x) = \arg\max_{k=1,\ldots,K} p_k(x)$ is the naive Bayesian classifier. In the following subsections, we delve into the intriguing phenomenon of "perfect classification", which is exclusively encountered in functional data classification, as elucidated in Section 5.1. Additionally, we thoroughly examine literature to shed light on the minimax optimal classifier in Section 5.2.

5.1 | Perfect classification phenomenon

Delaigle and Hall [2012] reported a unique phenomenon called “perfect classification”, wherein the misclassification error for binary classification can reach a value of zero. To be specific, they considered one-dimensional functional data $X(t)$ with domain $\mathcal{T} \subset \mathbb{R}$, such that $X(t)$ is generated from Π_1 with the prior probability π, and from Π_2 with the prior probability $1 - \pi$. The expectation operators $\{\mathrm{E}_k\}_{k=1,2}$ for the two classes are assumed to be $\mathrm{E}_1(X) = 0$ and $\mathrm{E}_2(X) = \mu(t)$, and the covariance function Ω of X is the same for both populations. To align the information of μ and the considered function space $\{\phi_j(t)\}$, μ is assumed to be decomposed as $\mu(t) = \sum_{j=1}^{\infty} \mu_j \phi_j(t)$. Within the established framework, Delaigle and Hall [2012] unveiled the perfect classification property for the sequence space

$$\Theta = \left\{ (\pi, \lambda_1, \mu_1, \lambda_2, \mu_2, \ldots) : \pi \in (0,1), \lambda_1 \geq \lambda_2 \geq \ldots, \sum_{j=1}^{\infty} \lambda_j < \infty, \sum_{j=1}^{\infty} \lambda_j^{-1} \mu_j^2 = \infty \right\}.$$

See Figure 5 for data visualization of both the perfect and imperfect classification examples. Given the intrinsically infinite dimensionality of functional data, it is conceivable that the minimal misclassification error within a specific function space can be reduced to zero. However, it is crucial to emphasize that achieving such an outcome in practice with finite dimensional data is highly impractical, except in exceedingly rare and anomalous situations that may be regarded as pathological. Consequently, this attribute underscores the substantial distinction between functional data and multivariate data. To address this problem, Delaigle and Hall [2012] initially proposed the centroid classifier. The benefit of the method lies in its sole reliance on the estimation of the mean and variance of the principal components, and it holds for both Gaussian and non-Gaussian functional data when Θ is taken into account. However, for Gaussian functional data, Delaigle and Hall [2012] claimed that the proposed centroid classifier is optimal only when $\sum_{j=1}^{\infty} \lambda_j^{-2} \mu_j^2 < \infty$, which resides in the “imperfect classification” situation. In another word, it cannot be assured that

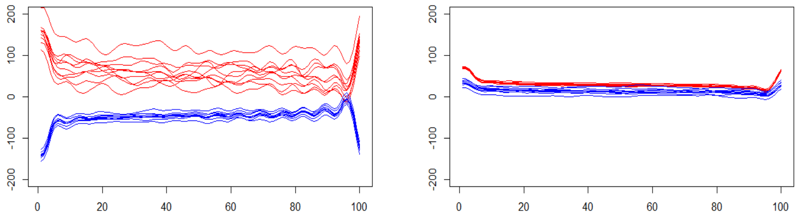

FIGURE 5 Illustration of both perfect and imperfect classification phenomenons. Left: perfect classification; Right: imperfect classification. Red curves: samples from the first class of functional data; Blue curves: samples from the second class of functional data.

the centroid classifier is optimal when confronted with the perfect classification scenario. For centroid classifier, when the condition for perfect classification is satisfied, it is uncertain if centroid classifier is optimal. When the condition of imperfect classification is met, it can be proved that the centroid classifier has the smallest misclassification risk among all classifiers.

Delaigle and Hall [2013] conducted further investigations into the phenomenon of perfect classification, specifically exploring scenarios where covariance functions exhibit variations. Berrendero et al. [2018] delivered similar results to validate the phenomenon in the view of RKHS, and they explained the vanishing Bayes risk by the mutual singularity of the two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ of the populations. Particularly, let $\mathbb{P}_k$ be the conditional distribution of $X|Y = k$, $k = 1, 2$. $\mathbb{P}_1$ and $\mathbb{P}_2$ are equivalent, denoted by $\mathbb{P}_1 \sim \mathbb{P}_2$, which indicates that $\mathbb{P}_1(B) = 0 \iff \mathbb{P}_2(B) = 0$, for any Borel set B; $\mathbb{P}_1$ and $\mathbb{P}_2$ are mutually singular, denoted by $\mathbb{P}_1 \perp\!\!\!\perp \mathbb{P}_2$, which indicates that there exists a Borel set B, such that $\mathbb{P}_1(B) = 1$ and $\mathbb{P}_2(B) = 0$. When two processes are Gaussian, Berrendero et al. [2018] aligned the perfect classification to mutual singularity and imperfect classification to measure equivalence, which provides an alternative viewpoint on the fundamental nature of the "perfect classification".

When the eigen-systems of the two classes have the common eigenfunctions, i.e. $\phi_j^{(1)} = \phi_j^{(2)}, j = 1, \ldots,$ with eigenvalues $\left\{\lambda_j^{(1)}\right\}_j$ and $\left\{\lambda_j^{(2)}\right\}_j$, respectively, the relationship of Π_1 and Π_2 is fully determined by the sequence $\left\{\left(\mu_j, \lambda_j^{(1)}, \lambda_j^{(2)}\right)\right\}_{j=1}^{\infty}$. In this degenerated scenario, Dai et al. [2017] demonstrated the necessary condition for the perfect classification in an explicit form, which is

$$\sum_{j=1}^{\infty} \mu_j^2/\lambda_j^{(2)} = \infty \quad \text{or} \quad \sum_{j=1}^{\infty} \left(\lambda_j^{(1)}/\lambda_j^{(2)} - 1\right)^2 = \infty. \tag{1}$$

It implies that perfect classification is only achievable when the two populations are sufficiently separated from each other in the sense that the infinite series that characterizes the distance between the mean functions or covariance functions is divergent. Furthermore, conditions in Equation (1) is not necessary for perfect classification. For example, it also happens when one class is Gaussian and the other is Lapalce, and both of them have the same parameters in Equation (1) (see Theorem 2 in Dai et al. [2017]).

While perfect classification offers the intriguing advantage of zero misclassification risk at the population level, the existing body of literature fails to address the rate of convergence for a specific classifier constructed from finite samples, such as those on the order of n. Consequently, evaluating the performance of various classifiers under perfect classification is often challenging. Furthermore, when contrasted with imperfect classification (where, for instance, both series converge), perfect classification represents a fundamentally simpler task, with the patterns of the two classes easily discernible in most practical scenarios.

5.2 | Minimax optimality

Generally, the efficacy of a classifier based on its risk convergence to the the Bayesian classifier is considered a fundamental metric. However, when two classifiers demonstrate similar convergence, relying solely on the measure of consistency becomes a challenging endeavor for determining the superior one. Therefore, employing additional criteria becomes imperative to make a more informed decision regarding the better classifier. Furthermore, when presented with a proficient classifier, one is invariably intrigued by its capacity to attain the optimal convergence rate that remains unattainable for any other classifier, except the Bayesian classifier. See Yang [1999], Mammen and Tsybakov [1999], Tsybakov [2004] for multivariate data classification.

Suppose $\mathcal{G}$ is the class of measurable functions, such that the naive Bayesian classifier $C^* \in \mathcal{G}$. Define the Minimax Excess Misclassification Risk (MEMR) as

$$\inf_{\widehat{C}_n} \sup_{C^* \in \mathcal{G}} E[R(\widehat{C}_n) - R(C^*)],$$

where the infimum is taken over all functional classifiers constructed using the training samples. MEMR offers a theoretical comprehension of the extent to which the Bayes risk can be approximated using finite training samples. Naturally, given any generic classifier $\widetilde{C}_n$, it has

$$\inf_{\widehat{C}_n} \sup_{C^* \in \mathcal{G}} E[R(\widehat{C}_n) - R(C^*)] \leq \sup_{C^* \in \mathcal{G}} E[R(\widetilde{C}_n) - R(C^*)],$$

such that the excess risk automatically bounds the MEMR from above. A rich body of literature has discussed the rates in different scenarios, see Biau et al. [2010], Baíllo and Cuesta-Albertos [2011], Meister [2016], Wang et al. [2023b,a]. Nonetheless, deriving the lower bound of convergence often proves challenging due to its exhaustive consideration of all potential classifiers. Any classifier is minimax optimal if and only if its excess risk matches the lower bound of MEMR. In the subsequent subsections, we highlight some selected work of functional data classification with regards to the minimax optimality across different function spaces.

5.2.1 | Gaussian functional data setting

Here we summarize the minimax optimality of FQDA and Functional Deep Neural Network (FDNN) proposed by Wang et al. [2023b], where they considered Gaussian process with the decomposition

$$X(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t), \quad \xi_j \sim N\left(\mu_{kj}, \lambda_j^{(k)}\right), \quad k = 1, 2,$$

under the k-th group and $t \subset \mathbb{R}^d$ with $d \geq 2$. Moreover, they explored the imperfect classification scenario, where both $\sum_{j=1}^{\infty} (\mu_{1j} - \mu_{2j})^2 / \lambda_j^{(2)}$ and $\sum_{j=1}^{\infty} \left(\lambda_j^{(1)} / \lambda_j^{(2)} - 1\right)^2$ are convergent. Besides, the square-integrable mean functions and bounded Ω_k result in another two convergent sequences: $\sum_{j=1}^{\infty} \mu_{kj}^2$ and $\sum_{j=1}^{\infty} \lambda_j^{(k)}$. In order to explicate the decay rate of the four sequences, the authors undertake an assessment using two distinct sequence spaces: the hyperrectangle and Sobolev ball. For the sake of clarity, we will concentrate solely on the Sobolev ball scenario, while acknowledging that the hyperrectangle method holds true in an analogous fashion. For any decay rate $\nu \in (0,1)$, by denoting Sobolev ball $S^{\nu} = \left\{\boldsymbol{a} = (a_1, a_2, \ldots) : \sum_{j=1}^{\infty} |a_j| j^{\nu} \leq M\right\}$ for some $M > 0$, Wang et al. [2023b] considered the parameter space $\Theta(\nu_1, \nu_2) := \{\, \boldsymbol{\theta} : \left\{\mu_{1j}^2 \vee \mu_{2j}^2\right\}_{j \geq 1} \in S^{\nu_1}, \left\{\lambda_j^{(1)} \vee \lambda_j^{(2)}\right\}_{j \geq 1} \in S^{\nu_1}, \left\{\left(\mu_{1j} - \mu_{2j}\right)^2 / \lambda_j^{(2)}\right\}_{j \geq 1} \in S^{\nu_2}, \left\{(\lambda_j^{(1)} / \lambda_j^{(2)} - 1)^2\right\}_{j \geq 1} \in S^{\nu_2} \}$, where ν_1 governs the separation of the two populations, and ν_2 governs the smoothness of the mean functions and covariance functions. By further assuming that $X_i^{(k)}(t_1), \ldots, X_i^{(k)}(t_M)$, $i = 1, \ldots, n_k, k = 1, 2$ are observed on evenly spaced $t_1, \ldots, t_M \in \mathcal{T}$, Wang et al. [2023b] showed that the convergence rate is in the order of n when the functional data are fully observed. Given that the complete trajectory of the data is observed, the contribution of smoothness is deemed to be negligible, thereby rendering the rate of separation between the two populations as the primary determining factor. For a finite M, a crucial threshold denoted by M^* can be identified, marking a phase transition point at which the convergence rate is equivalent to that of the fully observed functional data when the value of M exceeds M^*, with the rate then becoming solely dependent on the sample size n. Conversely, when M is less than M^*, the rate is entirely contingent upon the value of M itself, as $M^{-\nu_1 \nu_2/(1+\nu_2)}$.

In order to achieve the EMER, Wang et al. [2023b] found that FQDA, which utilizes the truncated quadratic discriminant function under the Gaussianity to approximate Bayesian classifier, is a minimax optimal approach for both fully observed and discretely observed data, provided that the truncation parameter is determined to be optimal. Moreover, they showed that the with a carefully selected network structure, FDNN is also minimax optimal up to a log factor. These findings highlight the theoretical exploration into the minimax optimality of functional classifiers. Nonetheless, the parameter under consideration lacks the necessary generality, and its extension to the perfect classification scenario is not feasible.

5.2.2 | Non-Gaussian functional data setting

Building upon the same framework, Wang et al. [2023a] undertook a more in-depth investigation into the realm of minimax optimality, extending beyond the scope of Gaussian functional data. In particular, they assume that the log-likelihood ratio belongs to a complex function space defined by Hölder smoothness, which extends from the parametric context in Wang et al. [2023b] to a non-parametric problem. For instance, Wang et al. [2023a] provided an example with student's t distribution, such that

$$X(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t), \quad \xi_j \sim t_{\nu_{kj}}, \quad k = 1, 2,$$

where ν_{kj} is the degree of freedom for the jth score and kth group. The log-likelihood function is thus written as $\sum_{j=1}^{\infty} \left\{\log e_j - \frac{\nu_{1j}+1}{2} \log\left(1 + \frac{\xi_j^2}{\nu_{1j}}\right) + \frac{\nu_{2j}+1}{2} \log\left(1 + \frac{\xi_j^2}{\nu_{2j}}\right)\right\}$, where e_j is some constant only depending on ν_{kj}. This discriminant function markedly differs from the Gaussian case. When contemplating a broad spectrum of discriminant functions, it becomes essential to approach them in a non-parametric manner. In particular, they established the EMER and validated the minimax optimality of the considered FDNN classifiers.

Different from Wang et al. [2023b] and Wang et al. [2023a], which describe the function space by defining the distribution of each principle component in its decomposed form, Meister [2016] adopted a holistic perspective and kept the full continuum of the function, imposing both smoothness conditions and complexity constraints on the considered function space. They have shown that by the general argument that the excess mass is bounded from above by the integrated squared regression risk. According to their theories, the proposed Nadaraya–Watson type classifier does not require knowledge of the smoothness degree parameter and, still, it leads to the optimal speed of convergence. Kraus and Stefanucci [2019] proposed regularized linear classifiers with domain selection and showed that possibly zero misclassification rate can be achieved despite that the training set consists of incomplete functions observed on different subsets of the domain.

6 | OPEN-SOURCE SOFTWARE

Unlike most machine learning areas that have abundant open-source software and data repositories, there has been relatively limited emphasis on promoting open FDA resources. As a consequence, comparing existing classifiers and evaluating new algorithms becomes challenging. Nevertheless, we make an earnest attempt to list publicly available R functions and packages in Table 1. In terms of computational cost, we evaluate both computational complexity and efficiency within a moderate sample size context. When compared to alternative methods, it is worth noting that DNN exhibits the highest model complexity and comparatively lower computational efficiency. Nevertheless, the utilization of dimensional reduction strategies as described in Wang et al. [2023a] and Wang and Cao [2023] substantially mitigates the computational burden, resulting in a more moderate level of computational cost.

TABLE 1 Prevalent `R` functions and packages for functional classifiers

method	function	package	platform	cost
kernel method [Ferraty and Vieu, 2006]	**classif.kernel**	`fda.usc`	CRAN	low
k–nearest neighbour [Ferraty and Vieu, 2006]	**classif.knn**	`fda.usc`	CRAN	low
generalized additive model [Ramsay and Silverman, 2005]	**classif.gsam**	`fda.usc`	CRAN	low
DD-Classifier [Mosler and Mozharovskyi, 2017]	**ddalphaf.classify**	`ddalpha`	CRAN	low
FDNN [Wang et al., 2023a]	**M_dnn.1d**, **M_dnn.2d**	–	GitHub	moderate
multi-class FDNN [Wang and Cao, 2023]	**mfdnn.1d** , **mfdnn.2d**	–	GitHub	moderate

7 | DISCUSSION

The main objective of our review is to develop an overarching understanding of existing functional data classification methods, emphasizing their similarities, differences, and connections with conventional multivariate classification approaches. We also introduce a new series classification methods under the framework of popular deep neural networks. We place emphasis on the works that are based on 2D and 3D functional data. Notably, we delve into a comprehensive discussion of the evaluation criteria employed in the functional data approach, particularly those that incorporate the increasingly popular DNN approach. We aim to help connect the machine learning, computer science and many other applied science communities with the challenges and opportunities in analyzing functional data. The investigation of classification problems can inspire other related research areas. For instance, results from functional data classification problems may benefit data collection and design. Given a pilot study with functional data that have known classes and may be either densely or sparsely observed, Li and Xiao [2020] identified the optimal sampling time points to collect observations for a new subject using linear discriminant analysis.

Despite progress being made in this field, existing methods often involve a single or finite number of random functions observed in the same domain. In practical applications, multivariate functional data are neither restricted to lie on the same interval nor to have one-dimensional domains. For example, Delaigle and Hall [2013] considered classification of functional data when the training curves are not observed on the same interval. Happ and Greven [2018] explored some complex functional data that consist of functions and images from a brain imaging dataset. The integration of these data types has opened up new research areas, specifically focusing on how classification can be effectively conducted on such data. Furthermore, as the number of random functions increases, both basis-based and FPCA-based classification algorithms face a significant increase in computational burden. Recently, Xue et al. [2023] tackled classification of high-dimensional functional data, in which each observation is potentially associated with a large number of functional processes. They proposed a penalized classifier and established discriminant set inclusion consistency in the sense that the classification responsible functional predictors include those of the underlying optimal classifier. Their work also sheds some light on the high-dimensional FDA from the classification point of view. All these developments suggest a promising research direction that seeks to address the challenges presented by high-dimensional and complex functional data, prompting the need for further research endeavors in this domain. Another promising future direction is the classification for dependent functional data Hörmann and Kokoszka [2010], Cao [2014], Cao and Wang [2018]. There are two popular types of dependent functional data structures: time series of curves and spatially distributed curves. Examples include daily curves of financial transaction data and daily patterns of geophysical and environmental data. Kokoszka [2012] reviewed recent research on dependent functional data. To the best of our knowledge, there is lack of literature on dependent functional data classification. Exploration in this research direction would benefit both FDA and classification domains.

references

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009) Functional logistic discrimination via regularized basis expansions. *Communications in Statistics. Theory and Methods*, **38**, 2944–2957.

Baíllo, A., C. and Cuesta-Albertos, J. A. (2011) Supervised classification for a family of gaussian functional models. *Scandinavian Journal of Statistics*, **38**, 480–498.

Berrendero, J. R., Cuevas, A. and Torrecilla, J. L. (2018) On the use of reproducing kernel hilbert spaces in functional classification. *Journal of the American Statistical Association*, **113**, 1210–1218.

Berrendero, R. J., Bueno-Larraz, B. and Cuevas, A. (2023) On functional logistic regression: some conceptual issues. *Test*, **32**, 321–349.

Beyaztas, U. and Shang, H. (2022) A robust functional partial least squares for scalar-on-multiple-function regression. *Journal of Chemometrics*, **36**, e3394.

Biau, G., Bunea, F. and Wegkamp, M. (2005) Functional classification in hilbert spaces. *IEEE Trans. Info. Theory*, **51**, 2163–2172.

Biau, G., Cérou, F. and Guyader, A. (2010) Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Trans. Info. Theory*, **56**, 2034–2040.

Blanquero, R., Carrizosa, E., Jiménez-Cordero, A. and Martín-Barragán, B. (2019) Variable selection in classification for multivariate functional data. *Information Sciences*, **481**, 445–462.

Bongiorno, G. E. and Goia, A. (2016) Classification methods for hilbert data based on surrogate density. *Computational Statistics and Data Analysis*, **99**, 204–222.

Cao, G. (2014) Simultaneous confidence bands for derivatives of dependent functional data. *Electronic Journal of Statistics*, **8**, 2639 – 2663.

Cao, G. and Wang, L. (2018) Simultaneous inference for the mean of repeated functional data. *Journal of Multivariate Analysis*, **165**, 279–295.

Cérou, F. and Guyader, A. (2006) Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, **10**, 340–355.

Chang, C., Chen, Y. and Ogden, R. T. (2014) Functional data classification: a wavelet approach. *Computational Statistics*, **29**, 1497–1513.

Chen, L.-H. and Jiang, C.-R. (2018) Sensible functional linear discriminant analysis. *Computational Statistics & Data Analysis*, **126**, 39–52.

Cuesta-Albertos, J. A., Fraiman, R. and Ransford, T. (2007) A sharp form of the Cramér-Wold theorem. *Journal of Theoretical Probability*, **20**, 201–209.

Cuevas, F., Febrero, M. and Fraiman, R. (2006) On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, **51**, 1063–1074.

Dai, W. and Genton, M. G. (2018) An outlyingness matrix for multivariate functional data classification. *Statistica Sinica*, **28**, 2435–2454.

— (2019) Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, **131**, 50–65.

Dai, X., Müller, H.-G. and Yao, F. (2017) Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, **104**, 545–560.

Darabi, N. and Hosseini-Nasab, S. M. E. (2020) Projection-based classification for functional data. *Statistics*, **54**, 544 – 558.

Delaigle, A. and Hall, P. (2012) Achieving near-perfect classification for functional data. *Journal of the Royal Statistical Society, Series B*, **74**, 267–286.

— (2013) Classification using censored functional data. *Journal of the American Statistical Association*, **108**, 1269–1283.

Ferraty, F. and Romain, Y. (2011) *The Oxford handbook of functional data analaysis*. Oxford University Press.

Ferraty, F. and Vieu, P. (2003) Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, **44**, 161–173.

— (2006) *Nonparametric Functional Data Analysis: Theory and Practice.* Berlin: Springer Series in Statistics, Springer.

Fisher, A. R. (1936) The use of multiple measurements in taxonomic problems. *Annuals of Eugenics*, **7**, 179–188.

Fraiman, R. and Muniz, G. (2001) Trimmed means for functional data. *Test*, **10**, 419–440.

Galeano, P., Joseph, E. and Lillo, R. E. (2015) The Mahalanobis distance for functional data with applications to classification. *Technometrics*, **57**, 281–291.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2011) Penalized functional regression. *Journal of Computational and Graphical Statistics*, **20**, 830–851.

Hall, P., Poskitt, D. S. and Presnell, B. (2001) A functional data-analytic approach to signal discrimination. *Technometrics*, **43**, 1–9.

Happ, C. and Greven, S. (2018) Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, **113**, 649–659.

Hlubinka, D., Gijbels, I., Omelka, M. and Nagy, S. (2015) Integrated data depth for smooth functions and its application in supervised classification. *Computational Statistics*, **30**, 1011–1031.

Hörmann, S. and Kokoszka, P. (2010) Weakly dependent functional data. *The Annals of Statistics*, **38**, 1845 – 1884. URL: https://doi.org/10.1214/09-AOS768.

Hsing, T. and Eubank, R. (2015) *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Hubert, M., Rousseeuw, P. J. and Segaert, P. (2015) Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, **11**, 445–466.

James, G. M. and Hastie, T. (2001) Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B*, **63**, 533–550.

Jiang, C.-R. and Chen, L.-H. (2020) Filtering-based approaches for functional data classification. *Wiley Interdisciplinary Reviews: Computational Statistics*, **12**, e1490.

Kokoszka, P. (2012) Dependent functional data. *International Scholarly Research Notices*, **2012**, 30 pages.

Kraus, D. and Stefanucci, M. (2019) Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, **106**, 161–180.

Kwon, A. M., Ouyang, M. and Cheng, A. Y. (2016) Resampling-based classification using depth for functional curves. *Communications in Statistics - Simulation and Computation*, **45**, 3329 – 3338.

Leng, X. and Müller, H. (2006) Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, **22**, 68–76.

Li, B. and Yu, Q. (2008) Classification of functional data: A segmentation approach. *Computational Statistics & Data Analysis*, **52**, 4790–4800.

Li, C. and Xiao, L. (2019) Optimal design for classification of functional data. *The Canadian Journal of Statistics*, **48**, 285–307.

— (2020) Optimal design for classification of functional data. *Canadian Journal of Statistics*, **48**, 285–307.

Li, X. and Ghosal, S. (2018) Bayesian classification of multiclass functional data. *Electronic Journal of Statistics*, **12**, 4669–4696.

López-Pintado, S. and Romo, J. (2005) Depth-based classification for functional data. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*.

López-Pintado, S., Sun, Y., Lin, J. K. and Genton, M. G. (2014) Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, **8**, 321–338.

Mammen, E. and Tsybakov, A. B. (1999) Smooth discrimination analysis. *The Annals of Statistics*, **27**, 1808–1829.

Meister, A. (2016) Optimal classification and nonparametric regression for functional data. *Bernoulli*, **22**, 1729–1744.

Mercer, J. (1909) Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, **209**, 415–446.

Moindjié, I.-A., Dabo-Niang, S. and Preda, C. (2022) Classification of multivariate functional data on different domains with partial least squares approaches. *arXiv preprint arXiv:2212.09145*.

Morris, S. J. (2015) Spline estimators for semi-functional linear model. *Annual Review of Statistics and Its Application*, **2**, 321–359.

Mosler, K. and Mozharovskyi, P. (2017) Fast dd-classification of functional data. *Statistical Papers*, **58**, 1055–1089.

Müller, H.-g. (2005) Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, **32**, 223–240.

Preda, C., Saporta, G. and Lévéder, C. (2007) PLS classification of functional data. *Computational Statistics*, **22**, 223–235.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis, Second Edition*. New York: Springer Series in Statistics.

Rossi, F. and Villab, N. (2006) Support vector machine for functional data classification. *Neurocomputing*, 730–742.

Sang, P., Kashlak, B. A. and Kong, L. (2022) A reproducing kernel hilbert space framework for functional classification. *Journal of Computational and Graphical Statistics*, 1–9.

Sguera, C., Galeano, P. and Lillo, R. (2014) Spatial depth-based classification for functional data. *TEST*, **23**, 725–750.

Shin, H. (2008) An extension of fisher's discriminant analysis for stochastic processes. *Journal of Multivariate Analysis*, **99**, 1191—-1216.

Torrecilla, J. L., Ramos-Carreno, C., Sanchez-Montanes, M. and Alberto, S. (2020) Optimal classification of gaussian processes in homo- and heteroscedastic settings. *Statistics and Computing*, **30**, 1091–1111.

Tsybakov, A. B. (2004) Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, **32**, 135–166.

Wang, J., Chiou, J. M. and Müller, H. G. (2016a) Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Wang, J.-L. W., Chiou, J.-M. and Müller, H.-G. (2016b) Review of functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Wang, S. and Cao, G. (2023) Multiclass classification for multidimensional functional data through deep neural networks. *arXiv:2305.13349*.

Wang, S., Cao, G. and Shang, Z. (2023a) Deep neural network classifier for multi-dimensional functional data. *Scandinavian Journal of Statistics*, in press.

Wang, S., Shang, Z., Cao, G. and Liu, S. J. (2023b) Optimal classification for functional data. *Statistica Sinica*, **34**.

Wu, Y. and Liu, Y. (2013) Functional robust support vector machines for sparse and irregular longitudinal data. *Journal of Computational and Graphical Statistics*, **22**, 379–395.

Xue, K., Yang, J. and Yao, F. (2023) Optimal linear discriminant analysis for high-dimensional functional data. *Journal of the American Statistical Association*.

Yang, Y. (1999) Minimax nonparametric classification. i. rates of convergence. ii. model selection for adaptation. *IEEE Transactions on Information Theory*, **45**, 2271–2292.

Yao, F., Wu, Y. and Zou, J. (2016) Probability-enhanced effective dimension reduction for classifying sparse functional data. *Test*, **25**, 1–22.

Yu, W., Wade, S., Bondell, D. H. and Azizi, L. (2022) Nonstationary gaussian process discriminant analysis with variable selection for high-dimensional functional data. *Journal of Computational and Graphical Statistics*, 1–13.

Zhang, Y.-C. and Sakhanenko, L. (2019) The naive bayes classifier for functional data. *Statistics & Probability Letters*, **152**, 137–146.

Zhu, H., Brown, P. J. and Morris, J. S. (2012) Robust classification of functional and quantitative image data using functional mixed models. *Biometrics*, **68**, 1260–1268.

Zhu, H., Vannucci, M. and Cox, D. D. (2010) A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics*, **00**, 463–473.