Received: 23 December 2023 | Revised: 29 May 2024 | Accepted: 3 July 2024

DOI: 10.1002/wics.70001

SYSTEMATIC REVIEW

WIREs COMPUTATIONAL STATISTICS

WILEY

# Functional data analysis using deep neural networks

**Shuoyang Wang**[1] | **Wanyu Zhang**[2] | **Guanqun Cao**[3] | **Yuan Huang**[2]

[1]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, Kentucky, USA

[2]Department of Biostatistics, Yale University, New Haven, Connecticut, USA

[3]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA

**Correspondence**
Shuoyang Wang, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA.
Email: shuoyang.wang@louisville.edu

**Funding information**
Division of Computer and Network Systems, Grant/Award Numbers: CNS-2319342, CNS-2319343

**Edited by:** James E. Gentle, Commissioning Editor and David W. Scott, Editor-in-Chief

**Abstract**

Functional data analysis is an evolving field focused on analyzing data that reveals insights into curves, surfaces, or entities within a continuous domain. This type of data is typically distinguished by the inherent dependence and smoothness observed within each data curve. Traditional functional data analysis approaches have predominantly relied on linear models, which, while foundational, often fall short in capturing the intricate, nonlinear relationships within the data. This paper seeks to bridge this gap by reviewing the integration of deep neural networks into functional data analysis. Deep neural networks present a transformative approach to navigating these complexities, excelling particularly in high-dimensional spaces and demonstrating unparalleled flexibility in managing diverse data constructs. This review aims to advance functional data regression, classification, and representation by integrating deep neural networks with functional data analysis, fostering a harmonious and synergistic union between these two fields. The remarkable ability of deep neural networks to adeptly navigate the intricate functional data highlights a wealth of opportunities for ongoing exploration and research across various interdisciplinary areas.

This article is categorized under:

Data: Types and Structure > Time Series, Stochastic Processes, and Functional Data

Statistical Learning and Exploratory Methods of the Data Sciences > Deep Learning

Statistical Learning and Exploratory Methods of the Data Sciences > Neural Networks

**KEYWORDS**

deep learning, functional data analysis, neural networks

## 1 | INTRODUCTION

### 1.1 | What is functional data analysis

Functional data analysis (FDA) is an evolving field that focuses on the analysis and theoretical exploration of data that offers insights into curves, surfaces, or any entities within a continuous domain. This continuous domain can span one dimension (1D), two dimensions (2D), or three dimensions (3D), encompassing aspects like a time span, a spectrum of

  

pixels or voxels, among other relevant scenarios. The inherent complexity of this field arises from the infinite-dimensional characteristic of the data.

In contrast to classical multivariate data, functional data is characterized by the inherent dependence and smoothness observed within each data curve. Rather than focusing on random variables, FDA involves examining random processes, which requires an implicit exploration of an infinite-dimensional function space. Over recent decades, FDA has garnered increasing attention within the statistical community, resulting in an extensive body of literature on the subject. For thorough understanding and additional insights, readers are encouraged to refer to the recent review articles (Morris, 2015; Salil & Staicu, 2023; Wang et al., 2016; Wang, Huang, & Cao, 2023).

There are three fundamental topics in FDA, which are functional data regression (Cardot, 2000; Rice & Wu, 2001), functional data classification (Araki et al., 2009; Shin, 2008), and functional data representation (Hall et al., 2006; Hall & Mohammad, 2006). Functional data regression encompasses a wide range of topics, such as exploring the relationships between functional responses and functional or scalar predictors, and reconstructing the underlying functions from functional observations. These tasks aim to uncover and understand the underlying patterns in data that evolve over a continuum. This form of analysis helps researchers make predictions and understand the dynamics of change across the continuous domain. Second, functional data classification focuses on categorizing functional observations into distinct groups or classes. This is especially useful in scenarios where the data can be naturally divided into categories, and there is interest in predicting the membership of new and unlabeled functional observation. Techniques used in this area often involve finding decision boundaries in the infinite-dimensional space or its finite counterpart that best separate different classes. Lastly, functional data representation seeks to efficiently identify the underlying structure of functional data. This often involves extracting features, reducing dimensionality, or finding a basis set that succinctly captures the essence of the data. Techniques such as functional principal component analysis (FPCA) are commonly employed to represent the data in a low rank space while preserving its essential characteristics.

## 1.2 | Why using deep neural networks?

Despite flourishing for over two decades, traditional approaches in FDA frequently depend on linear models to explore data structures. While these models serve as foundational tools, they pose substantial limitations by failing to capture the intricate and often nonlinear relationships inherent in functional data. Typical examples include functional linear regression model (Ramsay & Silverman, 2005), linear discriminant analysis for classification (Delaigle & Hall, 2012a), and FPCA (Happ & Greven, 2018; Ramsay & Silverman, 2005). Furthermore, the effective dimensionality reduction schemes intrinsic to conventional FDA methodologies typically yield a finite-dimensional representation of the data, an approach that, albeit necessary, can oversimplify or miss intricate data features. However, the advent of deep learning, particularly deep neural networks (DNNs), has ushered in a transformative approach to handling these complexities within FDA. As a popular approach with applications on approximating and estimating of multivariate functions, DNN is one of the most promising and vibrate areas in deep learning due to its learning capabilities of nonlinearities. In most recent years, DNNs have been applied in various nonparametric regression problems and shown to successfully overcome the curse of dimensionality in nonparametric regression. Examples include Schmidt-Hieber (2020); Bauer and Kohler (2019); Liu, Boukai, and Shang (2021); Liu, Shang, and Cheng (2021). Additionally, recent literature has demonstrated the significant success of DNNs in tackling classification problems with near optimal convergence rates Kim and Lim (2022); Bos and Schmidt-Hieber (2022).

As a powerful tool in dealing with high-dimensional data, DNNs are far superior to their traditional counterparts. They thrive in high-dimensional spaces, a characteristic that is particularly beneficial for functional data, often represented in many dimensions (Advani et al., 2020; Liu et al., 2017). The "curse of dimensionality" that plagues traditional models is less of an issue for DNNs, thanks to their ability to derive latent representations and unveil essential, albeit non-directly observable, structures within the data (Bauer & Kohler, 2019; Kim et al., 2021; Schmidt-Hieber, 2020). Moreover, deep learning models offer unparalleled flexibility. They can seamlessly cater to various data types and structures, whether the functional data are represented as curves, shapes, or images (Gal & Ghahramani, 2016; LeCun et al., 2015). This versatility in handling different data constructs is generally absent in more traditional statistical methods, which are often constrained by rigid structural assumptions.

The employment of DNNs in FDA also facilitates the integration of advanced neural network architectures. For instance, recurrent neural networks (RNNs) excels in contexts where temporal sequences are critical (Giles et al., 1994; Pearlmutter, 1989; Robinson, 1994). These sophisticated techniques enable a more profound extraction of spatial–

temporal features, often overlooked by conventional methods. In terms of predictive accuracy, DNNs have an edge due to their capacity for learning detailed, hierarchical representations of data (Ansuini et al., 2019; Raghu et al., 2019), significantly enhancing the precision of predictions—an essential factor in fields requiring reliable forecasts. Additionally, with the rise of computational power and parallel processing technologies, DNNs are uniquely scalable, capable of managing the large-scale datasets commonly encountered in contemporary functional data applications.

To date, there has been limited research on FDA using DNNs, with many studies focusing primarily on single-hidden-layer neural networks (Barron, 1993, 1994; White, 1990). Compared to DNNs, single-hidden-layer neural networks are simpler, limiting them primarily to more fundamental tasks. They can perform only a more basic level of feature extraction and are less capable in handling complex, nonlinear decision boundaries compared to their multi-layered counterparts. Despite this, we include these studies to our review, as they offer critical insights into applying neural networks to functional data regression, classification, and representation. Furthermore, they hold the potential for future extension into the deeper realms of deep learning, paving the way for more advanced applications and understanding in this field.

The structure of this paper is organized as follows. Section 2 presents various motivational examples encompassing traditional and next-generation functional data, as well as a compilation of open-source tools and software pertinent to functional data using DNNs. Preliminary notations, the dimension reduction framework for FDA, and the fundamentals of neural networks are introduced in Section 3. In Section 4, we explore the applications of functional data regression using neural networks across four distinct models, accompanied by an in-depth discussion of the evaluation criteria pertinent to each case. Section 5 pivots to the domain of classification, discussing the utilization of DNNs and shedding light on the relevant theoretical underpinnings reflected in current studies. Section 6 delves into the representation of functional data through the prism of DNNs. The concluding remarks and discussions in Section 7 encapsulate the paper, highlighting key findings and suggesting avenues for future exploration.

## 2 | FUNCTIONAL DATA AND ITS APPLICATIONS

This section is dedicated to showcasing the diverse and versatile applications of FDA across a range of fields. We explore a variety of examples, extending from traditional functional data contexts to cutting-edge, next-generation scenarios. The aim is to highlight the wide-ranging utility and adaptability of FDA techniques. In addition, we will introduce and discuss several open-source packages that are instrumental in integrating DNN into FDA challenges.

### 2.1 | Traditional functional data

#### 2.1.1 | Daily temperatures data

The Daily Temperatures Dataset (Ramsay & Silverman, 2005), a prominent resource in FDA, offers a comprehensive collection of daily temperature measurements, typically spanning several years. This dataset stands out for its representation of a continuous process—temperature over time—making it an ideal tool for exploring various aspects of seasonal trends, variations, and long-term climate patterns. One of the most referenced versions of this dataset is the Canadian Weather dataset, accessible through the `fda` package in `R`. It encompasses daily temperature and precipitation measurements from 35 Canadian weather stations over the course of a year. See Figure 1 for a representative display of temperature profiles from selected weather stations.

#### 2.1.2 | TIMIT speech data

The TIMIT speech data, sourced from the TIMIT Acoustic-Phonetic Continuous Speech Corpus, was meticulously compiled by the National Technical Information Service under the auspices of the United States Department of Commerce. This database holds a significant place in advancing the fields of speech recognition and functional data classification research. Key studies, such as those by Wang, Shang, et al. (2023); Rao and Reimherr (2023), have effectively utilized DNNs to build classifiers using this dataset. Within the rich repository of the TIMIT database, five distinct phonemes were carefully selected and transcribed: "sh" as in "she," "dcl" as in "dark," "iy" representing the vowel sound in "she,"

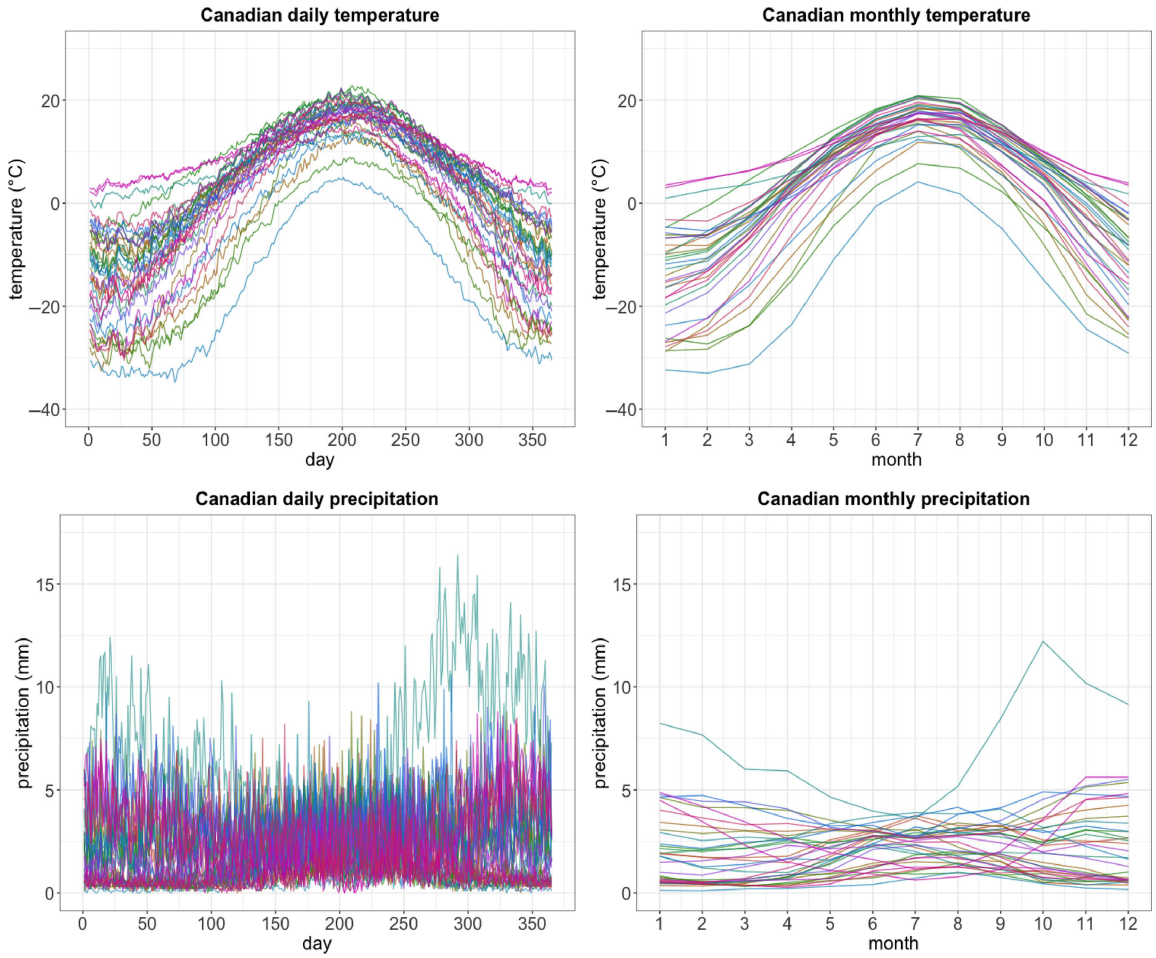


**FIGURE 1** Illustration of Daily Temperatures Dataset: Temperature and precipitation measurements from 35 Canadian weather stations. Left: Daily measurements. Right: Monthly measurements.

"aa" from the vowel in "dark," and "ao" from the first vowel in "water." Prior to any analytical processing, each speech frame was subjected to a log-periodogram transformation, making the data optimally suited for speech recognition applications. Typically, each frame in this dataset encompasses 400 samples, collected at a 16-kHz sampling rate, with the analysis primarily focusing on the top 150 frequencies. To visualize these variances, Figure 2 displays 30 log-periodograms from four different phoneme categories respectively, effectively highlighting the unique acoustic features of each selected phoneme.

## 2.2 | Next-generation functional data

### 2.2.1 | Fashion-MNIST dataset

This 2D functional data example is derived from the Fashion-MNIST dataset, accessible at https://github.com/zalandoresearch/fashion-mnist. Fashion-MNIST is an alternative to the traditional MNIST dataset, comprising 60,000 training images and 10,000 testing images. These images represent articles of clothing categorized into 10 distinct classes, including T-shirts/tops, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. Each image is meticulously normalized and centered within a $28 \times 28$ pixel frame, and features grayscale anti-aliased renderings of fashion products. The pixel values range from 0 to 255, denoting varying intensity levels, which naturally lends itself to a functional data interpretation. Each image can be viewed as a function mapped over a square domain. This characteristic facilitates the application of functional data methodologies in constructing classifiers based on these fashion item samples. Figure 3 depicts a selection of images from the Fashion-MNIST dataset, showcasing the diverse intensity patterns and clothing categories.

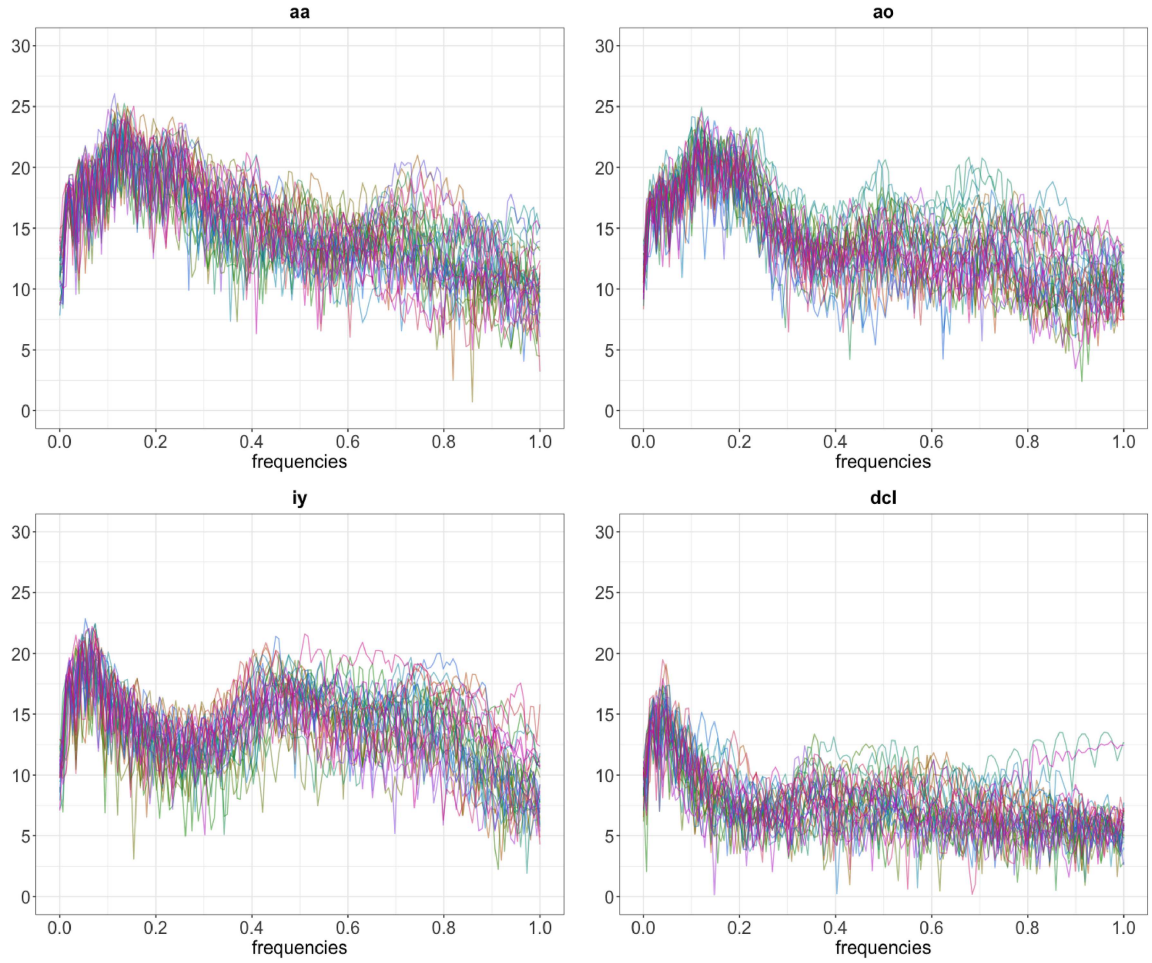


**FIGURE 2** Illustration of TIMIT speech data: A sample of 30 log-periodograms for each of the "aa," "ao," "iy," and "dcl" phonemes.

**FIGURE 3** Selected images from Fashion-MINIST dataset.

### 2.2.2 | Positron emission tomography scan dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database, found at adni.loni.usc.edu, is an invaluable repository for multidimensional functional data, particularly highlighted by its extensive collection of positron emission tomography (PET) scans. Originating from a comprehensive longitudinal multicenter study, ADNI plays a crucial role in advancing the identification and validation of key clinical, imaging, genetic, and biochemical biomarkers that are essential for monitoring the onset and progression of Alzheimer's disease (AD). The dataset presents several unique

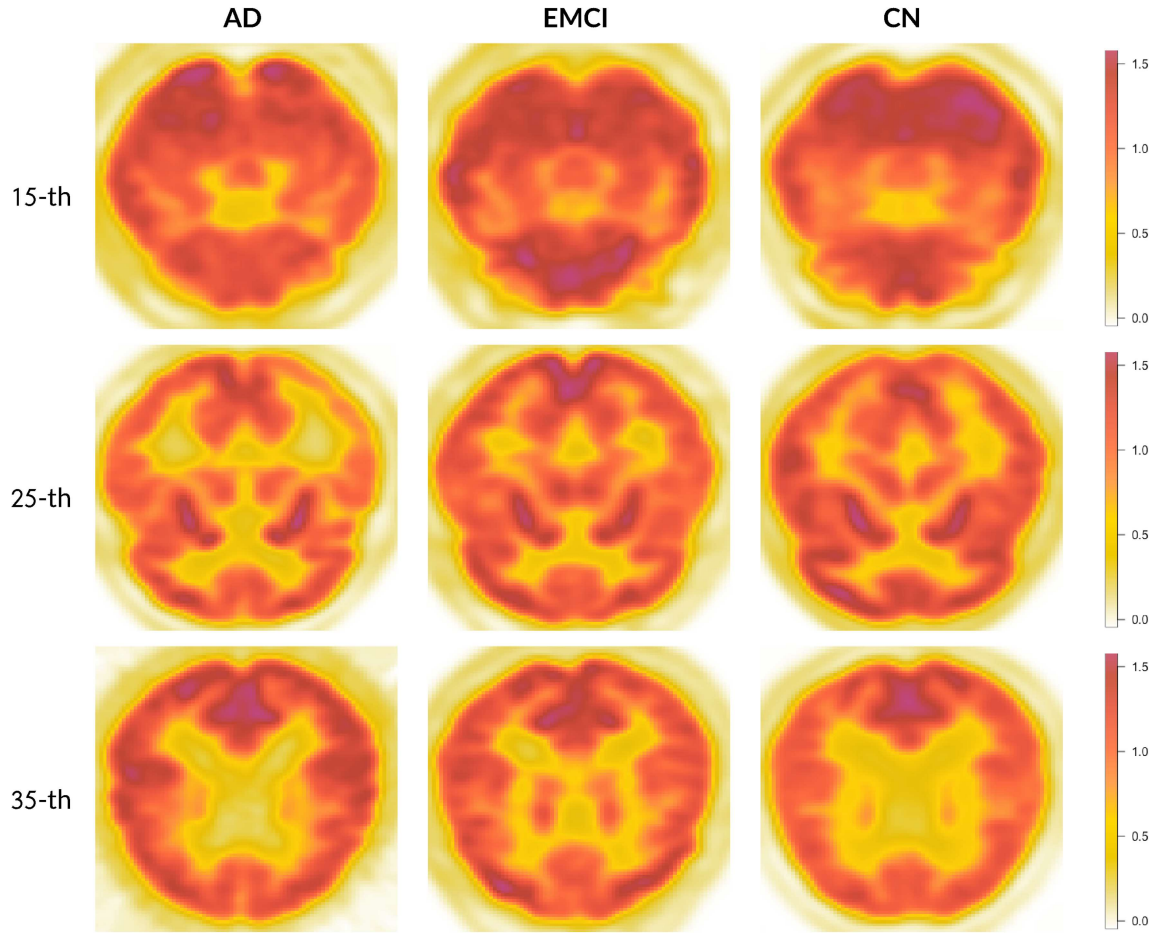


**FIGURE 4** Illustration of ANDI data. Selected images of the 15-th, the 25-th and the 35-th slices of AD group (left), EMCI group (middle), and CN group (right).

challenges. For instance, reconstructing brain scans from AD patient samples requires the use of advanced functional regression techniques to precisely capture the complex details in these scans. Moreover, effectively distinguishing between various stages of Alzheimer's disease—from control group (CN) subjects, to those with early mild cognitive impairment (EMCI), and patients with a formal AD diagnosis—necessitates the implementation of sophisticated binary or multiclass classification systems to analyze the extensive PET imaging data.

In Figure 4, the imaging data is meticulously organized, categorizing the scans by distinct groups and specific slices. Each image segment has undergone thorough spatial normalization and several post-processing steps to ensure the integrity and completeness of the data for detailed analysis and accurate classification. These processed scans adhere to a standardized format, each framed within dimensions of $79 \times 95 \times 68$ voxels, resulting in 68 2D image slices per patient. Each of these slices comprises $79 \times 95$ pixels, totaling $7,505$ discernible pixels per 2D slice. In a 3D analysis context, every complete brain scan encompasses a maximum of $510,340$ voxels, derived from the calculation $79 \times 95 \times 68$. This stringent standardization guarantees a uniform and all-encompassing dataset, laying the groundwork for sophisticated analytical methodologies. Addressing regression hurdles within both 2D and 3D ADNI data, Wang et al. (2021) and Wang and Cao (2022) have pioneered deep neural network-based approaches for FDA, representing substantial progress within this domain.

## 2.3 | Open-source software via neural networks

In the realm of machine learning, there exists a rich ecosystem of open-source software and extensive data resources. However, this abundance is not mirrored in the field of FDA, particularly in its integration with deep learning. The

FDA domain, especially in the context of deep learning applications, is still emerging and faces a notable shortage of specialized software solutions tailored for advanced analysis. This lack of resources presents challenges in comparing established methods and evaluating new algorithms. Despite these hurdles, our research includes a comprehensive compilation of available public functions and packages, which are detailed in Table 1. It is important to note that some of these resources primarily focus on single-hidden-layer neural networks, which might not fully align with the conventional definition of DNNs. Nevertheless, for completeness and to aid the readers, these tools are also included in our overview.

## 3 | PRELIMINARIES

Classical function data is a collection of independently and randomly observed curves which are real-valued functions (Wang et al., 2016). 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 $L_2(\mathcal{T})$, $\mathcal{T} \in \mathbb{R}$. Generally speaking, the next-generation functional data can be extensively denoted as $X(\mathbf{t})$ for some $\mathbf{t} \in \mathcal{T} \subset \mathbb{R}^d$, $d \geq 2$. 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.

For simplicity, we let $\mathcal{T} = [0,1]$, and denote our functional data in $L_2(\mathcal{T})$ without further explanation. To distinguish the data type, let $Y(t)$ be the functional response, and $Y$ be the scalar response. Similarly, let $X_1(t),...,X_p(t)$ be the functional covariates, and $X_1,...,X_p$ as the scalar covariates. Let $\epsilon(t) \in L_2(\mathcal{T})$ be the random error curve with mean zero, and $\epsilon$ be the random variable with mean zero and finite variance. Denote $\left\| \cdot \right\|$ as the $L_2$ norm.

### 3.1 | Neural networks

In the realm of machine learning, neural networks constitute a significant paradigm, capable of implementing complex and high-dimensional mappings from inputs to outputs. These mappings are achieved through layers of interconnected nodes or neurons, each responsible for processing input data, applying specific transformations, and forwarding the result. Mathematically, a typical feedforward neural network can be described as follows: Consider an input vector $\mathbf{x} \in \mathbb{R}^q$, the output of each neuron in the first hidden layer is calculated by applying a nonlinear activation function $\sigma(\cdot)$ to a linear transformation of the input. Specifically, if we denote the weight matrix connecting the input layer to the first hidden layer by $\mathbf{W}_1 \in \mathbb{R}^{N_1 \times q}$ and the bias vector by $\mathbf{b}_1 \in \mathbb{R}^{N_1}$, the output of the first hidden layer, $\mathbf{z}_1$, is given by

$$\mathbf{z}_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1).$$

This process continues through subsequent layers until the final layer is reached. The ultimate output of the network, $\mathbf{Y}$, for regression tasks, or a class label for classification tasks, is then obtained by possibly a different transformation of the final hidden layer. The expressiveness and flexibility of neural networks stem from their ability to approximate

**TABLE 1** Key software functions (with supporting software) on GitHub for FDA using neural networks.

| Article | Task | Function (software) |
|---|---|---|
| Wang et al. (2021) | Regression | `FDADNN (R)` |
| Yao et al. (2021) | Regression | `AdaFNN (Python)` |
| Wang and Cao (2022) | Regression | `RDNN (R)` |
| Wu et al. (2023) | Regression | `fosr_dp`, `fosr_dppm (Julia)` |
| Thind et al. (2023) | Regression | `FNN (R)` |
| Wang, Cao, and Shang (2023) | Classification | `M_dnn.1d`, `M_dnn.2d (R)` |
| Wang and Cao (2023b) | Classification | `mfdnn.1d`, `mfdnn.2d (R)` |
| Wang and Cao (2023a) | Data representation | `funnol (Python)` |

virtually any continuous function, as established by the universal approximation theorem. This quality makes them particularly suited to tasks where the underlying data distribution or true function.

As technology and computational sciences progressed, the advent of DNNs marked a transformative point in machine learning history. While traditional neural networks comprise only a couple of hidden layers, DNNs extend this architecture dramatically. A "deep" network is characterized by having multiple, often dozens of, hidden layers, each contributing to more refined and abstract representations of the input data. The formulation of a DNN can be viewed as a continuation of the previous neural network model, with additional layers compounding the complexity of transformations:

$$\mathbf{z}_{l+1} = h(\mathbf{W}_{l+1}\mathbf{z}_l + \mathbf{b}_{l+1}),$$

where $l$ represents the current layer, and $\mathbf{z}_l$ the output from the previous layer. This recursive transformation through each layer serves to build increasingly abstract features from the raw input. See Figure 5 for illustration.

The impetus behind the push for deeper networks is twofold. First, deep architectures allow for the hierarchical representation of data, reflecting the multifaceted structures found in real-world information. Complex concepts are often composed of simpler ones; deep networks can learn these hierarchies in an automated manner, creating intricate, layered representations. Second, DNNs have proven incredibly effective in practice, setting performance benchmarks across various domains, such as image recognition, natural language processing, and even complex board games. This efficacy is attributed to their capacity for high-dimensional function approximation, and, importantly, to their ability to disentangle the factors of variation in the input data through successive stages of nonlinear transformation.

## 3.2 | Traditional dimension reduction of functional observations

In contrast to classical multivariate data, the inherently infinite-dimensional nature of functional data necessitates the use of dimension reduction techniques for effective analysis. These techniques are crucial for a wide range of statistical tasks in FDA, including regression, classification, and data representation. Broadly, there are two primary approaches for handling functional data. The first approach involves applying dimension reduction techniques to the functional observations, converting them into a more manageable lower-dimensional form. This transformed data can then be processed using conventional multivariate classification methods. Techniques in this category include FPCA (Karhunen, 1946), basis expansion (Ramsay & Silverman, 2005), partial least squares (Delaigle & Hall, 2012b), among others. The second approach maintains the continuous nature of the functional data (Meister, 2016). Here, methods are applied directly to the unaltered functional data, although some might still incorporate dimension reduction depending on the technique used. This category includes distance-based methods and those that utilize the unique structure of functional spaces, such as methods based in Reproducing Kernel Hilbert Space (Shin & Lee, 2016).

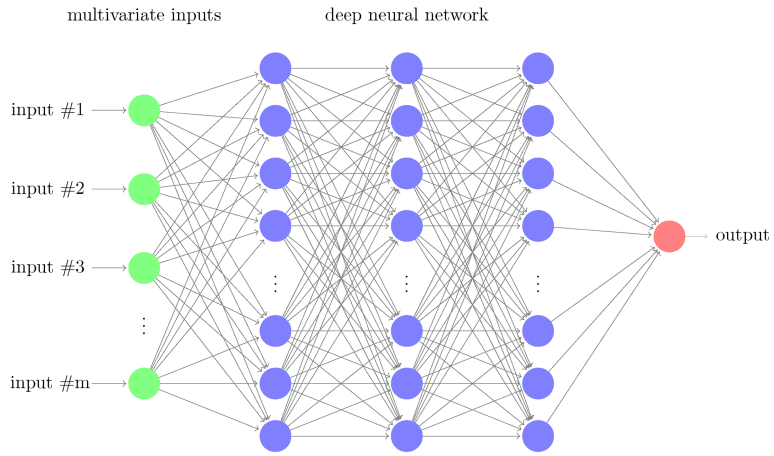


FIGURE 5 Visualization of DNN with three hidden layers and $m$-dimensional inputs.

Notably, FPCA stands as the most prominent and widely utilized method for dimension reduction in FDA, celebrated for its efficacy and clarity in interpretation (Leng & Müller, 2006; Müller, 2005). Consider a stochastic process $X(t)$, where $t \in \mathcal{T}$, with the integral $\int_{\mathcal{T}} \mathbb{E}\left[X(t)^2\right] dt$ being finite. This process $X(t)$ is characterized by an undetermined mean function $\mu_k(t)$ and an unknown covariance function $\Omega(t,t')$ for $t,t' \in \mathcal{T}$.

For a given set of $n$ random curves $\{X_1(t),...,X_n(t)\}$, the Karhunen–Loève expansion allows each $X_i(t)$ to be reformulated as:

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

where $\xi_{ij} = \int_{\mathcal{T}} (X_i(t) - \mu(t))\phi_j(t)dt$ are uncorrelated variables with mean 0 and variance $\lambda_j$, such that $\lambda_1 \geq \lambda_2 \geq ... \geq 0$, the $\phi_k(t)$ represents the corresponding orthonormal eigenfunction, such that $\int_{t \in \mathcal{T}} \phi_j(t)\phi_k(t)dt = \mathbb{I}(j=k)$, for $j,k \in \mathbb{N}^+$. These $\xi_{ij}$ are known as the Functional Principal Component Scores.

In practical applications, FPCA typically utilizes a truncated expansion that leverages either a predefined functional basis or a data-driven eigenbasis. This methodology affords enhanced adaptability and customization in the representation of functional data, enabling analyses that are both more detailed and accurate.

# 4 | FUNCTIONAL DATA REGRESSION

## 4.1 | Formulation of functional data regression

Regarding the functional data regression, researchers encounter a diverse array of tasks, each presenting unique challenges and complexities. These tasks range from reconstructing trajectories from functional observations to various forms of regression analyses, including scalar-on-function, function-on-scalar, and function-on-function regression. Historically, the bulk of scholarly literature in this field has concentrated on linear models, which, while foundational, often fall short in capturing the nuanced dynamics within the data. However, a recent surge in innovative research has seen the application of DNNs come to the fore, providing new methodologies to unravel the intricate nonlinear relationships inherent in these regression models. DNNs offer a significant leap forward, delivering more nuanced, adaptive, and accurate modeling capabilities that are particularly suited to managing the complexity of functional data.

### 4.1.1 | Trajectory recovery

In the realm of FDA, the estimation of mean functions stands as the foundational step, critical to subsequent analysis. For instance, when analyzing PET imaging data from Alzheimer's Disease patients, estimating the mean function of this data enables researchers to discern common patterns and features within the patient cohort. This process involves the precise determination of central tendencies within the infinite-dimensional functional data, serving as a basis for further statistical exploration and interpretation in various complex FDA problems. Specifically, for multidimensional functional data $Y(\mathbf{t})$, $\mathbf{t} \in \mathcal{T}^d \subset \mathbb{R}^d$, it follows that

$$Y(\mathbf{t}) = f(\mathbf{t}) + \eta(\mathbf{t}) + \epsilon(\mathbf{t}),$$

where $f$ is a map from $\mathcal{T}^d$ to $\mathbb{R}$, $\eta \in L_2\left(\mathcal{T}^d\right)$ is an empirical process with mean zero. Conventionally, the literature primarily focus on one-dimensional case, where $d = 1$ (Cardot, 2000). When transitioning to higher dimensions with $d \geq 2$, the complexity of the function $f$ escalates. This rise not only hampers the efficiency of the estimation procedures, making them more computationally intensive and challenging to navigate, but also introduces substantial obstacles in the formulation of a uniform estimator for $f$. This multidimensional context demands more sophisticated analytical tools and methodologies, as the intricacies of the functional behavior become increasingly difficult to approximate accurately and consistently across various dimensions of the data space.

In response to these challenges, Wang et al. (2021) spearheaded an innovative approach by employing DNNs for FDA. In their methodology, the observed grid points of functional observations are utilized as inputs, enabling the

recovered function to achieve the classical convergence rate characteristic of nonparametric regression. This technique marked a significant departure from traditional methods, demonstrating the efficacy of deep learning in handling the intricate structures inherent in functional data.

Building on this foundational work, Wang and Cao (2022) further advanced the field by extending the application from Gaussian functional data to non-Gaussian contexts. This expansion enhanced the robustness of the estimator, allowing for more flexibility and reliability, particularly in scenarios where data do not adhere to the Gaussian assumption. By accommodating a broader range of data distributions, this refined approach underscores the adaptability and potency of DNNs in navigating the complex landscape of FDA.

### 4.1.2 | Scalar-on-function

As one of the most prominent regression models, scalar-on-function (SOF) regression enjoys widespread use across various fields, owing to its critical role in predicting a scalar outcome based on functional predictors, a scenario common in numerous practical applications. This methodology is perfectly illustrated by the analysis of the Tecator dataset (https://lib.stat.cmu.edu/datasets/tecator), where spectrometric curves, discretized across a grid of 100 wavelengths, serve as the functional predictors. In this scenario, the scalar response variable is the fat content in meat samples, determined through analytical chemical processing. This approach allows for a nuanced understanding of how the spectral data, representing various absorbance levels at different wavelengths, can effectively predict the fat composition in the meat, a vital factor in food quality and processing.

In the realm of traditional linear and generalized linear models, several prominent methodologies are extensively employed, such as the functional linear model (Eilers et al., 2009):

$$Y = \beta_0 + \sum_{j=1}^{p} \int_{\mathcal{T}} \beta_j(t) X_j(t) dt + \epsilon,$$

the single-index model Ramsay and Silverman (2005):

$$Y = g\left( \sum_{j=1}^{p} \int_{\mathcal{T}} \beta_j(t) X_j(t) dt \right) + \epsilon,$$

among others. However, the definite structure in parametric fashion lacks the flexibility to accommodate various types of functional inputs. To generalize the regression model, consider

$$Y = \mathcal{A}\big(X_1(t), ..., X_p(t)\big) + \epsilon, \tag{1}$$

where $\mathcal{A}$ is a map from $L_2(\mathcal{T})^p$ to $\mathbb{R}$.

To address the estimation of $\mathcal{A}$, Delannay et al. (2004) propose the pioneer work, where they consider the functional radial basis function networks, and the estimator is represented by the single-layer neural network:

$$\widehat{\mathcal{A}} = \sum_{\ell=1}^{N} u_\ell \sigma_\ell \left( v_\ell^{-1} \left( \int_{\mathcal{T}} \omega(t)(X_1(t) - \mu_\ell(t) dt) \right)^{1/2} \right),$$

where $\sigma_\ell$ are the radial basis functions, $u_\ell$ are random weights, $v_\ell$ are scaling factors adjusted with the learning algorithm, $\omega(t)$ is the weight function, and $\mu_\ell(t)$ are centered functions adjusted by $X_1(t)$. The trivial choice of $\omega(t)$ is $\omega(t) = 1$, $X(t)$ and $\mu_\ell(t)$ are generally represented by same basis functions.

Rossi et al. (2005) employ the integral to the functional inputs, such that for each $X_j$, the first layer of neural network is calculated by

$$\sum_{\ell}^{N} u_\ell \sigma_\ell \left( v_\ell + \int_{\mathcal{T}} g_j(t) X_j(t) dt \right),$$

where $\sigma$ is the activation function, $g_j$ are coefficient functions represented by some topological basis, and $X_j$ are represented by some chosen basis functions. This methodology facilitates the transformation of functional inputs into scalar variables, which are then efficiently processed through subsequent layers of a multi-layer neural network. This nuanced integration preserves the intricate characteristics of the functional data while leveraging the advanced pattern recognition capabilities of DNNs. However, when representing the functional inputs $X_j$, they do not take advantage of the key information contained in the response $Y$ during its dimension reduction stage, and the choice of basis functions are subjective. Similar strategy can be found in Thind et al. (2023) and the functional basis neural network in Rao and Reimherr (2023).

To address the issue of basis function, Yao et al. (2021) propose the Adaptive Functional Neural Network (AdaFNN), which is an adaptive approach to find the optimal bases that utilizes the information on $Y$. Specifically, they additionally model the basis function $g_j$ by the neural network, such that $g_j(t) = \sigma_L(...\sigma_1(u_\ell t + v_\ell))$, where $L$ is the number of layers. Besides, to promote the representation of diverse and uncorrelated information about the function through different nodes in basis layer, a regularization strategy can be employed to maintain their orthogonality. This approach ensures that each node captures unique features of the function, enhancing the overall capacity and effectiveness of the model. Rao and Reimherr (2023) propose the similar functional direct neural network with regularization and develop functional gradient based optimization algorithm to optimize the network parameters.

In essence, Rossi et al. (2005), Yao et al. (2021), Thind et al. (2023) and Rao and Reimherr (2023) utilize the functional linear model as the initial layer in their methodologies. All subsequent operations within their frameworks build upon the foundational structure provided by the functional linear model, ensuring a coherent integration with traditional FDA paradigms. Comparably, Wu et al. (2023) propose the utilization of projection scores $\{\xi_{kj}\}_{k=1}^{r_j}$ for the functional covariates $X_j$, achieved through various dimension reduction techniques. These finite-dimensional projection scores are then employed as inputs for the neural network. This approach effectively transforms the network into a conventional feedforward neural network tailored for multivariate data, aligning with standard practices in the field. It is important to highlight that when extracting the projection scores, the methodology employed involves adopting least squares estimation, particularly applied to the discretely observed functional covariates. The choice of basis includes the pre-fixed Fourier basis or the basis by the FPCA.

### 4.1.3 | Function-on-scalar

Function-on-scalar (FOS) regression is a statistical technique used when the response variable is a function and the predictors are scalar variables. This approach is particularly useful for analyzing data where the response is inherently functional over a domain, such as time or space, and the predictors are fixed values. Using this method, the relationship between fractional anisotropy curves, derived along the midsagittal skeleton of the corpus callosum, and a range of scalar covariates—including gender, age, Alzheimer's disease status, mild cognitive impairment status, and so forth—was investigated. The primary objective was to understand how these scalar variables influence the fractional anisotropy curves, which are integral to assessing the integrity of white matter tracts. Conventionally, the linear FOS regression model (Faraway, 1997; Ramsay & Silverman, 2005) is given by

$$Y(t) = \beta_0(t) + \sum_{j=1}^{p} X_j \beta_j(t) + \epsilon(t).$$

Luo and Qi (2023b) further explore the nonlinear FOS for the one-dimensional functional data. Specifically, they consider a general map $\mathcal{F}$ from $\mathbb{R}^p$ to $L_2(\mathcal{T})$, and use $\mathcal{F}_t$ to denote the value of $\mathcal{F}$ at $t$. Therefore, the linear model can be considered as a trivial case of

$$Y(t) = \mathcal{F}_t(X_1, ..., X_p) + \epsilon(t). \tag{2}$$

Some other applications of Model (2) include the functional additive mixed models (Scheipl et al., 2015).

To estimate $\mathcal{F}_t$, they propose the least-squares objective function with smoothness penalty, and the estimation is via single-hidden-layer functional neural network, such that $\widehat{\mathcal{F}}_t = \int_{\mathcal{T}} \delta(t,x)\sigma\left(\gamma_0(t,x) + \sum_{j=1}^{p} X_j\gamma_j(t,x)\right)dx$, where $\delta$ and $\{\gamma\}_{j=0}^{p}$ are smooth bivariate functions on $\mathcal{T}\times\mathcal{T}$, and they are represented by the tensor product of basis functions. $\sigma$ is an activation function which is bounded and Lipschitz continuous. Note that the popular ReLU activation function in the classical universal approximation theory is not qualified in the framework of function-valued maps. This is because the ReLU function is unbounded, unlike other activation functions such as sigmoid function, the hyperbolic tangent function, and the Gaussian function, which violates the crucial assumption of the functional universal approximation theory. In particular, Model (2) can also be applied to the imaging response, which was considered by (Zhang et al., 2023) using DNN approach.

#### 4.1.4 | Function-on-function

When the analysis involves both response variables and covariates represented as functional data, it is categorized under the function-on-function (FOF) regression model. This approach is widely applied across various domains, one notable example being the analysis of the Capital Bike Share System in Washington, DC (Cao et al., 2020). In this study, FOF regression was utilized to model hourly bike pick-up counts as a functional response, with observed hourly humidity and wind speed serving as functional predictors. The primary objective of this analysis was to forecasting hourly bike rental volumes and taking into account weather conditions and whether it was a working day. Such insights are crucial for effective trip planning and the efficient management of the bike-sharing system. There are a few existing models for FOF regression, including the linear FOF regression model:

$$Y(t) = \beta_0(t) + \int_{\mathcal{T}} X_1(s)\beta(s,t)ds + \epsilon(t),$$

or the linear concurrent model:

$$Y(t) = \beta_0(t) + X_1(t)\beta(t) + \epsilon(t).$$

See Ramsay and Silverman (2005) for more details.

To explore a flexible relationship between the functional response and functional covariates, Luo and Qi (2023a) consider general nonlinear FOF regression model, where

$$Y(t) = \mathcal{H}_t(X_1) + \epsilon(t), \tag{3}$$

where $\mathcal{H}_t$ is the value of a map $\mathcal{H}$ at $t$, and map $\mathcal{H}$ is from $\mathcal{T}$ to $\mathcal{T}$. To estimate $\mathcal{H}_t$, they propose the least-squares objective function with smoothness penalty, and the estimation is via single-hidden-layer functional neural network, such that $\widehat{\mathcal{H}}_t = \int_{\mathcal{T}} \delta(t,x)\sigma\big(\gamma_0(t,x) + \int_{\mathcal{T}} X_1(s)\gamma_1(t,x,s)\big)dx$, where $\delta,\gamma_0 \in \mathcal{T}\times\mathcal{T}$ and $\gamma_1 \in \mathcal{T}\times\mathcal{T}\times\mathcal{T}$ are infinitely differentiable functions. Similar to Luo and Qi (2023b), the bivariate functions $\delta,\gamma_0$ are approximated by tensor product basis functions, and the trivariate function is depicted through basis functions with reduced dimensions to enhance computational efficiency.

### 4.2 | Evaluation of functional regression

To evaluate the regression estimators, approximation error and estimation error are two fundamental concepts, each reflecting distinct aspects of the analytical process. On the one hand, approximation error essentially reflects the gap between the true function and the best possible prediction we can achieve within the chosen hypothesis space or model family. Estimation error, on the other hand, emerges from the randomness in the data sample utilized for training the model. Given that we almost never have access to the entire population of data, we rely on samples to train our models. In practice, both types of errors coexist, and their interplay is a central focus in developing statistical models.

For example, the approximation error is generally associated with the concept of bias in the bias-variance tradeoff, where a more complex model (such as neural network space) might have less approximation error (or bias). On the

contrary, estimation error is related to the variance component, highlighting the variability that comes from the data's randomness and the learning algorithm's sensitivity to this randomness. In the realm of DNNs, the estimation error tends to be substantial due to the sheer number of parameters involved, making its computation intricate owing to the system's overparameterization and the indefinite characteristics of the estimators involved. These factors pose significant challenges to advanced theoretical methodologies, as the unconventional architecture of these networks goes beyond the traditional statistical frameworks typically applied to more straightforward, well-defined forms.

In the existing literature, most of the work provide the universal approximation error of the proposed algorithms. The universal approximation theorem asserts that a neural network has the capability to approximate any desired function to a certain degree of accuracy, given carefully selected parameters. For the single-hidden-layer neural network, Thind et al. (2023) provide the universal approximation theory for SOF on hypercube, which has the same dimension as the sample size; Luo and Qi (2023a) and Luo and Qi (2023b) deliver similar functional universal approximation theorems for functional outputs, where the nonlinear map under FOS and FOF regression model can be approximated by some functional neural network with one hidden layer, and the error is arbitrarily small. For neural networks with multiple layers, Yao et al. (2021) show the universal approximation theory when the integral of functional components is well approximated.

In comparison, there has been limited research on estimation error of DNNs, particularly in providing non-asymptotic convergence rates as a function of sample size. Wang et al. (2021) and Wang and Cao (2022) have established the convergence rate of the DNN estimator in empirical norms, achieving a minimax rate of convergence comparable to that presented in Stone (1982), up to a logarithmic factor. Leveraging insights from the deep learning field, their derived convergence rate remains uninfluenced by the dimensionality of the functional inputs. Furthermore, they present a data-driven neural network structure along with its non-asymptotic rate, offering valuable guidance on choosing an optimal structure to expedite convergence in practical applications.

## 5 | FUNCTIONAL DATA CLASSIFICATION

There is less literature on functional data classification compared to functional data regression. However, this domain holds significant potential, as illustrated by datasets like the Berkeley Growth Study and the TIMIT speech data. The Berkeley Growth Study provides functional data through growth curves, tracking children's heights at different ages, which can be used to distinguish gender (Rao & Reimherr, 2023). On the other hand, the TIMIT dataset, sourced from the TIMIT Acoustic-Phonetic Continuous Speech Corpus, offers a wealth of speech signals as functional data. This dataset is employed for classifying various phonemes, such as "sh" in "she" and "dcl" in "dark," demonstrating the diverse applications of functional data classification (Wang, Shang, et al., 2023).

While certain methodologies from Section 4, as illustrated by Yao et al. (2021); Thind et al. (2023); Rao and Reimherr (2023), can be seamlessly adapted to classification tasks by incorporating a link function, and their proposed network structures remain applicable, it is important to note that even though these methods offer broad applicability across varied tasks, they might not be the most optimal solution for specific problems. Thus, our primary emphasis is on literature explicitly addressing functional data classification, which substantiates the optimality of the functional classifier. To elucidate the existing literature on functional classifiers, we begin by thoroughly outlining the constructions of these classifiers in Section 5.1. Subsequently, we provide the evaluations in Section 5.2.

### 5.1 | Deep neural network classifier for functional data

For $K \geq 2$, suppose we examine a $K$-class classification problem, wherein the functional observations are defined over the space $\mathcal{X}$. Let $(X_i, Y_i)$, $i = 1, ..., n$, be i.i.d. random pairs of observations, where $X_i \in \mathcal{X}$ and $Y_i \in \{1, ..., K\}$. For a new observation $X$, the classification task is to predict the class label $Y$ by a classifier $C : \mathcal{X} \to \{1, ..., K\}$, based on finite sample $\{(X_i, Y_i)\}_{i=1}^n$. More specifically, the classification rule based on finite sample is defined as $\widehat{C}_n \equiv \widehat{C}_n((X_1, Y_1), ..., (X_n, Y_n))$. We denote the prior probability $\pi_k = \mathbb{P}(Y = k)$, and the posterior probabilities $p_k(x) = \mathbb{P}(Y = k | X = x)$, $k = 1, ..., K$.

To address the fundamental binary classification task, Wang, Shang, et al. (2023) and Wang, Cao, and Shang (2023) propose the Functional Deep Neural Network (FDNN) method, where they assume that the functional observations admit the general decomposition

$$X(\cdot)=\sum_{j=1}^{\infty}\xi_j\phi_{jk}(\cdot),\quad k=1,2,$$

where $\left\{\phi_{j1}\right\}_{j=1}^{\infty}$ and $\left\{\phi_{j2}\right\}_{j=1}^{\infty}$ are the orthonormal basis of some function spaces, and $\{\xi_j\}_{j=1}^{\infty}$ represent random functional projection scores (FPS), each following a unique distribution that varies given different $k$ values. Particularly, FDNN first obtains the estimation of basis functions $\left\{\widehat{\phi}_{jk}\right\}$ by performing Karhunen–Loéve decomposition for the sample covariance functions $\widehat{\Omega}_k$. Subsequently, for some integer $J$, they extract the truncated $J$ FPS from the sample data function $X_i$ by integration, as

$$X_i(\cdot)\approx\sum_{j=1}^{J}\widehat{\xi}_{ij}\widehat{\phi}_{jk}(\cdot),\quad i=1,\ldots,n.$$

Denote $\widehat{\boldsymbol{\xi}}_J^{(i)}=\left(\widehat{\xi}_{i1},\ldots,\widehat{\xi}_{iJ}\right)$, and the training sample set $\{(X_i,Y_i)\}_{i=1}^{n}$ is reformulated as $\left\{\left(\widehat{\boldsymbol{\xi}}_J^{(i)},Y_i\right)\right\}_{i=1}^{n}$. Let the FDNN discriminant function be

$$\widehat{g}(\cdot)=\arg\min_{f\in\mathcal{D}}\frac{1}{n}\sum_{i=1}^{n}\max\left(1-g\left(\widehat{\boldsymbol{\xi}}_J^{(i)}\right)Y_i,0\right),$$

which minimizes the hinge loss function, and $\mathcal{D}$ is some DNN class with data-driven structure. Specifically, the selection of the class $\mathcal{D}$ is contingent upon the complexity inherent in the Bayes discriminant rule, as exemplified by the analysis of Gaussian functional data in Wang, Shang, et al. (2023) and non-Gaussian functional data in Wang, Cao, and Shang (2023). Based upon $\widehat{g}$, the FDNN classifier is constructed by

$$\widehat{C}^{FDNN}(X_i)=\begin{cases}1, & \widehat{g}\left(\boldsymbol{\xi}_J^{(i)}\right)\geq 0,\\ -1, & \widehat{g}\left(\boldsymbol{\xi}_J^{(i)}\right)<0.\end{cases}$$

To practically select the optimal structure of $\mathcal{D}$, including the truncation parameter $J$, Wang, Cao, and Shang (2023) recommend a data-splitting method through validation techniques.

## 5.2 | Evaluation of functional classifiers

When evaluating a classifier's performance, two primary benchmarks are commonly used. The first involves assessing the misclassification risk, defined as

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

which represents the probability of incorrect classifications, given collected samples. This measure offers a direct and intuitive evaluation of the classifier's accuracy. The second benchmark examines the classifier's consistency with the Bayes classifier. This approach is considered highly reliable, as it evaluates how closely a classifier approximates the optimal classification strategy. Specifically, a classifier is deemed consistent if its misclassification risk converges to the Bayesian risk, expressed as

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

where $C^*(x)=\arg\max_{\mathrm{k}=1,\ldots,\mathrm{K}}\mathrm{p}_{\mathrm{k}}(\mathrm{x})$ denotes the naive Bayesian classifier.

The effectiveness of a classifier is often measured by its risk convergence toward the Bayesian classifier, a critical benchmark in classifier assessment. However, distinguishing between two classifiers that exhibit similar convergence rates poses a challenge, as consistency alone may not suffice to identify the superior option. In such cases, incorporating additional evaluation criteria is essential for a more discerning comparison. Moreover, the discovery of a proficient classifier naturally raises interest in its ability to achieve the optimal convergence rate, a feat matched only by the Bayesian classifier. This aspect of classifier performance, especially in the context of multivariate data classification, has been explored in depth by Yang (1999); Mammen and Tsybakov (1999); Tsybakov (2004).

Suppose $\mathcal{G}$ represents a class of measurable functions that includes the naive Bayesian classifier $C^*$, and Minimax Excess Misclassification Risk (MEMR) is defined as:

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

where the infimum encompasses all functional classifiers constructed from the training samples. This definition of MEMR offers a theoretical framework to gauge the approximation to the Bayes risk using the finite training samples. For any general classifier $\tilde{C}_n$, it holds that:

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

indicating that the excess risk inherently sets an upper bound for MEMR. However, identifying the lower bound of convergence requires consideration of all possible classifiers. In the context of DNNs classifiers, minimax optimality is achieved if and only if the classifier's excess risk matches the lower bound of MEMR, thereby establishing it as the optimal choice among all classifiers.

To summarize, Wang, Shang, et al. (2023) demonstrate that to achieve EMER, the FDNN is optimally minimax (accounting for a logarithmic factor relative to sample size) in situations involving both fully and discretely observed data, provided that the truncation parameter is optimally chosen. Their work significantly advances the theoretical understanding of minimax optimality in functional classifiers, though it is noted that the study's parameters may not be broad enough to apply in scenarios requiring perfect classification. Building upon these findings, Wang, Cao, and Shang (2023) explore deeper into the realm of minimax optimality, moving beyond Gaussian functional data. They suggest that the log-likelihood ratio belongs to a complex function space characterized by Hölder smoothness, transitioning from the parametric framework of Wang, Shang, et al. (2023) to a more comprehensive nonparametric approach. Their proposed methods not only reaffirms EMER but also validates the minimax optimality of FDNN classifiers in this broader context. These two pivotal discoveries underscore the capability of DNN classifiers to be the optimal choice in highly intricate function spaces, thus positioning them as a formidable tool for tackling complex classification tasks in diverse functional data scenarios.

## 6 | FUNCTIONAL DATA REPRESENTATION

As mentioned in Section 3.2, FPCA has been considered as the most classical approach to represent functional data. See Ramsay and Silverman (2005); Hall and Mohammad (2006); Hall et al. (2006) for univariate FPCA and Happ and Greven (2018) for multivariate FPCA. However, as a linear projection approach, the FPCA have several drawbacks in representing functional data with complicated structures. Primarily, this method fails to capture the nonlinear structures inherent in functional data, often not providing any more effective information than the original observations. It also struggles with multivariate functional data, resorting to inadequate linear components when forced to combine all dimensions. Furthermore, FPCA assumes a common covariance function across observations, an assumption that, if incorrect, compromises the entire functional principal space, especially when data contain specific labels for individual trajectories.

In response to the inherent limitations of linear projections previously discussed, Wang and Cao (2023a) propose the Functional Nonlinear Learning approach. This methodology is inspired by the concept of an autoencoder, which was first applied to functional data by Hsieh et al. (2021). When the functional data is observed discretely across $M$

points, the data collection can be represented as $\{\mathbf{x}_m\}_{m=1}^{M}$, where $\mathbf{x}_m = \big(X_1(t_m),...,X_p(t_m)\big)$. The procedure they introduced is summarized as follows:

- Step 1 (encoding). Model the recurrent neural network (RNN) as

$$\mathbf{f}_m = \sigma(\mathbf{W}\mathbf{x}_m + \mathbf{U}\mathbf{f}_{m-1}),$$

and the latent feature $\mathbf{z}$ is represented by $\mathbf{z} = \sigma(\mathbf{V}\mathbf{f}_M)$.

- Step 2 (decoding). Model the recurrent neural network as

$$\tilde{\mathbf{f}}_m = \sigma\Big(\tilde{\mathbf{W}}\mathbf{z} + \tilde{\mathbf{U}}\tilde{\mathbf{f}}_{m-1}\Big),$$

and the reconstruction $\mathbf{x}_m$ for each $m$ is represented by $\mathbf{x}_m = \sigma\Big(\tilde{\mathbf{V}}\tilde{\mathbf{f}}_M\Big)$.

- Step 3 (supervised learning if necessary). Model the response variable $Y$ by the latent feature $\mathbf{z}$ with a single-layer neural network, such that $Y = \sigma(\mathbf{H}\mathbf{z})$.

By incorporating the RNN, the proposed method ensures continuity in the observed trajectories, making it more suitable for functional observations.

In conclusion, it is important to note that while the proposed method is not strictly a deep neural network approach due to its use of only a single-hidden-layer, it is still relevant to include here for its potential insights. The field of functional data representation through deep learning is indeed an emerging and rapidly evolving area. Currently in its early stages, it demands more thorough research and attention to fully explore its capabilities and broader implications in various applications.

## 7 | DISCUSSION

In summary, we have strived to offer an exhaustive understanding of the evolving landscape of functional data methodologies that leverage (deep) neural networks. Our exploration underscores not only their groundbreaking concepts but also their multifaceted applications, spanning regression, classification, and data representation. While we shed light on various methodologies, a critical component of our review has been the evaluation and comparison of these methods, ensuring readers obtain a panoramic view of the prevailing literature on this critical subject. Rather than restricting our scope to the conventional realm of 1D functional data, we have broadened our lens to encompass research grounded in 2D and 3D functional data as well. One of our core aspirations has been to serve as a bridge, connecting the domains of machine learning, computer science, and other applied sciences, spotlighting the myriad challenges and opportunities intrinsic to FDA. The fusion of DNNs into this field signifies a momentous shift away from the orthodox linear methodologies. By tapping into the nonlinear modeling prowess, adaptability, and superior feature extraction capabilities inherent to DNNs, we posit that researchers and practitioners are better positioned to extract profound insights from functional data, thereby catalyzing innovation across a spectrum of scientific fields.

Despite advancements in this field, there remains ample scope for further progression. For instance, several studies discussed in this review, such as Luo and Qi (2023b), are limited to neural networks with a single-hidden-layer. An exciting prospect lies in the exploration of multi-layer neural networks and the incorporation of multiple functional covariates, potentially facilitating intricate functional relationships. Furthermore, the potential applications of deep learning extend well beyond their current use. Areas ripe for exploration include functional data clustering and the analysis of functional time series, where the advanced capabilities of DNNs could uncover new insights and foster deeper understanding. The recent work by Ma et al. (2024) sheds lights on using DNN to handle functional time series.

Moreover, there is a notable absence of deep learning methodologies tailored to high-dimensional functional data, where the selection of critical functional features is paramount. A recent endeavor by Xue et al. (2023) addresses the classification of such data, where each observation is potentially linked with numerous functional processes. They introduced a penalized classifier and established discriminant set inclusion consistency, ensuring that the classification-responsible functional predictors encompass those of the optimal underlying classifier. This pioneering effort offers insights into high-dimensional FDA, charting possible future trajectories for both regression and classification tasks involving functional data using DNNs. Collectively, these evolutions underscore the burgeoning potential in addressing the intricacies of high-dimensional and intricate functional data, amplifying the call for intensified research in this sphere.

Compared to traditional FDA methods, DNN-integrated FDA approaches require more advanced computational skills. As with all deep learning approaches, computational time may increase due to the large number of tuning parameters and training strategies. This generally varies with the depth and width of the neural network, as well as with training epochs, batch sizes, algorithms, and other factors. In terms of model interpretability, although DNN-based methods provide a more flexible structure for addressing the problem, they lack the interpretability of classical statistical models. For example, in a functional linear model, the coefficient function represents the contribution of the corresponding covariate to the response variable at each point within the domain, offering clear interpretability. This level of interpretability is not evident in Equation (1). Therefore, it is recommended to apply existing DNN-based approaches when prediction is the primary focus. A future direction in this field may involve combining both interpretability and predictive ability, such as through semi-parametric modeling (Wang & Huang, 2024; Zhong & Wang, 2023).

## AUTHOR CONTRIBUTIONS

**Shuoyang Wang:** Data curation (equal); methodology (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Wanyu Zhang:** Data curation (equal); methodology (equal); visualization (equal); writing – review and editing (equal). **Guanqun Cao:** Methodology (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Yuan Huang:** Methodology (equal); validation (equal); writing – review and editing (equal).

## FUNDING INFORMATION

Cao's research was also partially supported by National Science Foundation under Grants CNS-2319342 and CNS-2319343.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

*Shuoyang Wang* https://orcid.org/0000-0001-6963-6267
*Guanqun Cao* https://orcid.org/0000-0002-6454-3210

## RELATED WIREs ARTICLES

On semiparametric regression in functional data analysis
Review on functional data classification
Model-based clustering and classification of functional data

## REFERENCES

Advani, S. M., Saxe, M. A., & Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, *132*, 428–446.

Ansuini, A., Laio, A., Macke, J. H., & Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, *32*.

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

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*, 930–945.

Barron, A. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, *14*, 115–133.

Bauer, B., & Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, *47*, 2261–2285.

Bos, T., & Schmidt-Hieber, J. (2022). Convergence rates of deep relu networks for multiclass classification. *Electronic Journal of Statistics*, *16*, 2724–2773.

Cao, G., Wang, S., & Wang, L. (2020). Estimation and inference for functional linear regression models with partially varying regression coefficients. *Stat*, *9*, e286.

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, *12*, 503–538.

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

Delaigle, A., & Hall, P. (2012b). Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, *40*, 322–352.

Delannay, N., Rossi, F., Conan-Guez, B., & Verleysen, M. (2004). Functional radial basis function networks (FRBFN). In *Proceedings of the 12th European Symposium on Artificial Neural Networks (ESANN 2004)* (pp. 313–318). https://perso.uclouvain.be/michel.verleysen/papers/esann04nd.pdf

Eilers, P. H., Li, B., & Marx, B. D. (2009). Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, *96*, 196–202.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, *39*, 254–261.

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. *Advances in Neural Information Processing Systems*, *29*, 1019–1027.

Giles, C. L., Kuhn, G. M., & Williams, R. (1994). Dynamic recurrent neural networks: Theory and applications. *IEEE Transactions on Neural Networks*, *5*, 153–156.

Hall, P., & Mohammad, H.-N. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society, Series B*, *68*, 109–126.

Hall, P., Müller, H. G., & Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, *34*, 1493–1517.

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

Hsieh, T.-Y., Sun, Y., Wang, S., & Honavary, V. (2021). Functional autoencoders for functional data representation learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)* (pp. 1–11). Society for Industrial and Applied Mathematics.

Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae. Series A*, *1*(34), 1–7.

Kim, H., & Lim, Y. (2022). Bootstrap aggregated classification for sparse functional data. *Journal of Applied Statistics*, *49*, 2052–2063.

Kim, Y., Ohn, I., & Kim, D. (2021). Fast convergence rates of deep neural networks for classification. *Neural Networks*, *138*, 179–197.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

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

Liu, B., Wei, Y., Zhang, Y., & Yang, Q. (2017). Deep neural networks for high dimension, low sample size data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. AAAI Press.

Liu, R., Boukai, B., & Shang, Z. (2021). Optimal nonparametric inference via deep neural network. *Journal of Mathematical Analysis and Applications*, *505*, 125561.

Liu, R., Shang, Z., & Cheng, G. (2021). *On deep instrumental variables estimate*. arXiv:2004.14954.

Luo, R., & Qi, X. (2023a). General nonlinear function-on-function regression via functional universal approximation. *Journal of Computational and Graphical Statistics*, *33*, 1–10.

Luo, R., & Qi, X. (2023b). Nonlinear function-on-scalar regression via functional universal approximation. *Biometrics*, *79*, 3319–3331.

Ma, T., Yao, F., & Zhou, Z. (2024). Network-level traffic flow prediction: Functional time series vs. functional neural network approach. *The Annals of Applied Statistics*, *18*, 424–444.

Mammen, E., & 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.

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

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

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*, 263–269.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Dickstein, J. S. (2019). On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 2847–2854). PLMR.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer Series in Statistics.

Rao, A. R., & Reimherr, M. (2023). Nonlinear functional modeling using neural networks. *Journal of Computational and Graphical Statistics*, *32*, 1–10.

Rice, J., & Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, *57*, 253–259.

Robinson, A. (1994). An application of recurrent neural nets to phone probability estimation. *IEEE Transactions on Neural Networks*, *5*, 298–305.

Rossi, F., Delannay, N., Conan-Guez, B., & Verleysen, M. (2005). Representation of functional data in neural networks. *Neurocomputing*, *64*, 183–210.

Salil, K., & Staicu, A.-M. (2023). Second-generation functional data. *Annual Review of Statistics and its Application*, *10*, 547–572.

Scheipl, F., Staicu, A.-M., & Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, *24*, 477–501.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, *48*, 1875–1897.

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

Shin, H., & Lee, S. (2016). An RKHS approach to robust functional linear regression. *Statistica Sinica*, *26*, 255–272.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, *10*, 1040–1053.

Thind, B., Multani, K., & Cao, J. (2023). Deep learning with functional inputs. *Journal of Computational and Graphical Statistics*, *32*, 171–180.

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

Wang, H., & Cao, J. (2023a). Functional nonlinear learning. *Journal of Computational and Graphical Statistics*, *33*, 1–11.

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

Wang, S., & Cao, G. (2022). Robust deep neural network estimation for multi-dimensional functional data. *Electronic Journal of Statistics*, *16*, 6461–6488.

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

Wang, S., Cao, G., & Shang, Z. (2021). Estimation of the mean function of functional data via deep neural networks. *Stat*, *10*, e393.

Wang, S., Cao, G., & Shang, Z. (2023). Deep neural network classifier for multi-dimensional functional data. *Scandinavian Journal of Statistics*, *50*, 1667–1686.

Wang, S., & Huang, Y. (2024). Dp2lm: Leveraging deep learning approach for estimation and hypothesis testing on mediation effects with high-dimensional mediators and complex confounders. *Biostatistics*, *25*, 818–832.

Wang, S., Huang, Y., & Cao, G. (2023). Review on functional data classification. *WIREs Computational Statistics*, *16*, e1638.

Wang, S., Shang, Z., Cao, G., & Liu, S. J. (2023). Optimal classification for functional data. *Statistica Sinica*, *34*, 1545–1564.

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, *3*, 535–550.

Wu, S., Beaulac, C., & Cao, J. (2023). Neural networks for scalar input and functional output. *Statistics and Computing*, *33*, 118.

Xue, K., Yang, J., & Yao, F. (2023). Optimal linear discriminant analysis for high-dimensional functional data. *Journal of the American Statistical Association*, *119*, 1055–1064.

Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. II. Model selection for adaptation. *IEEE Transactions on Information Theory*, *45*, 2271–2292.

Yao, J., Müller, J., & Wang, J.-L. (2021). Deep learning for functional data analysis with adaptive basis layers. In *Proceedings of the 38th International Conference on Machine Learning, Vol. 139 of Proceedings of Machine Learning Research* (pp. 11898–11908). PMLR.

Zhang, D., Li, L., Sripada, C., & Kang, J. (2023). Image response regression via deep neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*, 1589–1614.

Zhong, Q., & Wang, J.-L. (2023). Neural networks for partially linear quantile regression. *Journal of Business & Economic Statistics*, *42*, 603–614.

**How to cite this article:** Wang, S., Zhang, W., Cao, G., & Huang, Y. (2024). Functional data analysis using deep neural networks. *WIREs Computational Statistics*, *16*(4), e70001. https://doi.org/10.1002/wics.70001