Supervised and Ensemble Classification of Multivariate Functional Data: Applications to Lupus Diagnosis

by

Robert Buscaglia

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2018 by the
Graduate Supervisory Committee:

Yiannis Kamarianakis, Chair
Dieter Armbruster
Nicholas Lanchier
Robert McCulloch
Mark Reiser

ARIZONA STATE UNIVERSITY

August 2018

ABSTRACT

This dissertation investigates the classification of systemic lupus erythematosus (SLE) in the presence of non-SLE alternatives, while developing novel curve classification methodologies with wide ranging applications. Functional data representations of plasma thermogram measurements and the corresponding derivative curves provide predictors yet to be investigated for SLE identification. Functional nonparametric classifiers form a methodological basis, which is used herein to develop a) the family of ESFuNC segment-wise curve classification algorithms and b) per-pixel ensembles based on logistic regression and fused-LASSO. The proposed methods achieve test set accuracy rates as high as 94.3%, while returning information about regions of the temperature domain that are critical for population discrimination. The undertaken analyses suggest that derivate-based information contributes significantly in improved classification performance relative to recently published studies on SLE plasma thermograms.

To my beautiful family, for your love and support.

ACKNOWLEDGMENTS

It was an incredible adventure making such a large career change to mathematics and statistics. Obtaining a second PhD and a position as a university professor are lifelong goals that have been achieved through hard work and guidance from many individuals at Arizona State University. I want to thank all of the instructors and advisors that helped me to finish a second doctorate program. I want to especially express gratitude and thanks to Dr. Yiannis Kamarianakis, who guided me into my journey studying statistics, a field in which I am passionate and excited to pursue as a career. His exceptional patience and overwhelming determination to produce quality and meaningful research have shaped this dissertation. I am humbled by the extent of the education I have received.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

BSS	Best segment selection
CES	Combined ensemble strategy
DSC	Differential scanning calorimetry
ESFuNC	Ensemble of functional nonparametric classifiers
ENET	Elastic-net
FDA	Functional data analysis
FDO	Functional data object
FDSA	Functional segment discriminant analysis
FGLM	Functional generalized linear models
FGAM	Functional generalized additive models
FGKAM	Functional generalized kernel additive models
FGSAM	Functional generalized spectral additive models
FLR	Functional logistic regression
FPC	Functional principal component
FPCA	Functional principal component analysis
FSS	Forward segment selection
FuNC	Functional nonparametric classifiers
GAM	Generalized additive models
GCV	Generalized cross-validation
GES	Greedy ensemble strategy

GLM	Generalized linear models
HES	Hierarchical ensemble strategy
KCV	K-fold cross-validation
KNN	K-nearest neighbors
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LSA	Lease squares approximation
LOOCV	Leave-one-out cross-validation
LR	Logistic regression
ML	Maximum likelihood
NC	Nonparametric classification
PCA	Principal component analysis
PPC	Per-pixel classifier
PPE	Per-pixel ensemble
PW	Parzen window
QDA	Quadratic discriminant analysis
SLE	Systemic lupus erythematosus
VIF	Variance inflation factor
WKNN	Weighted-KNN

Chapter 1

INTRODUCTION

1.1 Overview

Modern scientific inquiry has reached a limit where mathematical treatment is a necessity to achieving promising results. A majority of the core sciences are now motivated by mathematical or statistical investigation, with fields such as big-data driving the way for novel and innovative findings. This work represents a synergistic journey of a classically trained experimental biochemist in an endeavor to improve advanced biophysical topics with modern statistical approaches.

One of the main priorities for preparing a second dissertation was to advance several of the published projects derived from biochemistry laboratories of previous employment. This goal was achieved through projects focused on technicality that can be used across disciplines. Several mathematical and statistical fields were explored in addition to the work developed herein. Studies in dynamical systems and perturbation methods provided extensions to previous kinetic experimentation. Numerical analysis and computational statistics have broadened potential impacts of nonlinear regression models, which have been used to provide thermodynamic profiles of complex unfolding mechanisms (Buscaglia et al. 2013).

A primary data set of interest used throughout this dissertation is the classification of disease states based on thermal denaturation of human plasma samples (Garbett et al. 2009; Garbett et al. 2007a; Garbett et al. 2007b). This diagnostic technique, known as plasma thermograms, has been under development as a means of investigating a variety of ailments such as autoimmune disorders, cancers, and diabetes. The biophysical experimentation required to collect plasma thermograms relates closely to the initial dissertation work developed by the author. Thermodynamic deconvolution of plasma thermograms is possible, but due to the density of proteins within the plasma proteome, it is inadequate to base models solely on thermodynamic properties. Instead, the resulting experimental signatures can be used as predictors in the identification of disease states.

The expectation for plasma thermograms is to support current clinically relevant diagnostic techniques with a fast and inexpensive means of evaluating difficult to diagnose ailments. One paramount plasma thermogram investigation is the classification of systemic lupus erythematosus (SLE) against non-SLE patients (Fish et al. 2010; Garbett et al. 2008). SLE is a difficult and spurious disease to identify correctly, with misdiagnosis occurring frequently with the potential for serious consequences (Narain et al. 2004). Plasma thermograms offer an additional means of identifying disease states that can improve a patient's diagnosis with minimal invasion. As a diagnostic technique, plasma thermograms have suffered from a lack of statistical treatment capable of clearly identifying disease states, a necessity for the technique to become clinically relevant.

To improve upon modern classification methods, it is necessary to combine aspects from several statistical learning methodologies. Statistical and machine learning

fields, including topics from functional data analysis (FDA), nonparametric classification (NC), supervised learning, and ensemble learning, are blended to provide a cutting edge analysis of SLE plasma thermograms. This motivates a deep investigation of the classification methodologies, highlighting potential pitfalls and encouraging improvements to modern techniques.

1.2 Statement of Research Topics

This work sets out to improve the analysis and understanding of plasma thermograms, with the chief focus of identifying SLE cases with high specificity and sensitivity. This dissertation can be summarized into three major research themes:

1. Application of modern statistical learning methodologies for the analysis and classification of SLE plasma thermograms.
2. Development of the ESFuNC curve classification algorithm and computational design for multivariate functional data.
3. Investigation of per-pixel classifiers and ensembles based on estimated class probabilities.

Chapter 2 gives the background of all major methodologies used in this work, along with an in-depth description of plasma thermograms. Chapter 3 evaluates SLE plasma thermograms from the standpoint of FDA. Functional representations of thermogram signatures are evaluated for improving disease identification. Chapter 4 presents supervised classification performance of nine modern statistical classifiers. Functional data classifiers and ensemble learning strategies are presented in Chapter 5. Chapter 6 implements a new computational approach to curve classification using

segmented classifiers. Multivariate functional ensemble strategies are developed, with empirical classification performance demonstrated through simulation. Additional ensemble methodologies are investigated in Chapter 7. Learning algorithms are developed that allow for optimization of the number of FPCs used in the estimation of the classifiers. Per-pixel classifiers (PPCs) are also examined and logistic regression (LR) is employed to construct ensembles of PPCs. Fused-LASSO is introduced as an alternative LR estimator, capable of producing ensembles while performing predictor selection and smoothing of predictor coefficients. This method provides both accurate classification and identification of regions of the functional domain that are important for population discrimination.

1.3 Results

The dissertation demonstrates that ensembles produced from segmented functional nonparametric classifiers (FuNC) are capable of improving the overall accuracy of SLE disease classification using only data from plasma thermograms. This work evaluates a variety of modern methodologies for empirical comparisons of classification performance. Although modern methodologies provide effective models, new approaches for the analysis of functional data are developed, which improve classification accuracy.

Functional data representations provide a flexible framework for evaluating classification performance based on multivariate supervised learning. Ensemble learning identifies classifier combinations that optimize classification accuracy. Stepwise

strategies are investigated, along with several approaches to how ensembles combining multivariate classifiers can be produced. The greedy ensemble strategy produces high classification accuracy but at computational expense. The combined ensemble strategy simplifies the computational burden while also simplifying the complexity of segmentation patterns and final ensembles. The hierarchical ensemble strategy implements dependence on how multivariate supervised classification models are produced and provides a compromise between effective classifiers and computational efficiency. These strategies are used to produce classifiers based on multivariate functional data to attain improved classification performance of SLE plasma thermograms beyond contemporary methods and recent literature studies.

Ensemble classifiers are then interrogated in an attempt to further improve classification performance. FPCA-based classifiers result in improvements relative to contemporary methods. Specifically, ensembles of FPCA-based classifiers are capable of producing accuracy rates equivalent to ESFuNC. Computational strategies for producing predictor sets based on PPCs are also introduced; PPCs produce a large set of predictors based on LOOCV estimated class probabilities. Potential pruning of the predictor set is evaluated using a variance inflation factor (VIF) based stepwise procedure. Additional ensemble strategies based on LR estimation are evaluated, and produce competitive methods that are improve in terms of accuracy in comparison with using traditional predictors.

Chapter 2

BACKGROUND

2.1 Supervised Learning

The main objective of the dissertation is to improve the classification of unknown data entries, which will be referred to as test set or out-of-set data. In supervised learning, one of the most widely accepted learning methods, a model is estimated through the use of labeled training data that consists of predictors paired with a known response. In the binary case, a model is produced to estimate if a given data object falls into Class 0 or Class 1. The labeled training data aids in the preparation of a classification model with accurate prediction of the known classes. The model is then tested with the goal of optimizing the out-of-set classification accuracy (Kotsiantis et al. 2007).

Supervised learning strategies evaluate a wide range of predictive models while performing parameter tuning in an attempt to maximize classification performance. Labeled training data is organized such that the predictors, or feature sets, are pruned such that the resulting feature subset does not contain irrelevant or redundant features (Yu and Liu 2004). Feature selection is a widely discussed literature topic with several methods that are now commonly employed in supervised learning schemes (Saeys et al. 2007). By pruning the feature space, the dimensionality of supervised learning algorithms is reduced and computational needs can be alleviated.

Data sets are typically given as a completed set of results, which are partitioned to create the labeled training data and out-of-set testing data (Kotsiantis et al. 2007). Commonly used partitioning schemes include two-third/one-third splits, where two-thirds of the data are used as a training set, and the remaining one-third of the data set used to test out-of-set generalization. More generalized forms exist such as K-fold cross-validation (KCV), which partitions the original data set into k-distinct sets. Training is then performed using $k-1$ of the data sets, and validation performed on the excluded set. When k is set to 3, this returns the two-third/one-third split. If k is set to be the sample size, this partitioning leads to leave-one-out cross-validation (LOOCV), where each data entry is considered as an out-of-set sample and how well the algorithm generalizes iteratively excluding each point can be evaluated. Computational considerations are required when deciding what partition sizes to use, with k set to 5 or 10 being commonly employed. Throughout this work, stratified KCV will be employed that provides the benefits of KCV while also ensuring that equal proportions of each class are represented within each fold.

Numerous supervised classification algorithms have been proposed in the literature (Hastie et al. 2009; Kotsiantis et al. 2007), with common modern examples including generalized linear models (GLM), generalized additive models (GAM), linear discriminant analysis (LDA), and NC. Each algorithm has potential benefits and different computational requirements. A survey of several common methods will be evaluated in Chapter 4, with applications of the different model estimation algorithms being applied to the classification of SLE plasma thermograms. The choice of algorithm is weighted based on empirical performance. When evaluating a new learning problem,

it is a common practice to evaluate several learning algorithms and choose the algorithm whose performance generalizes best to out-of-set results.

Three common metrics for classification performance are accuracy, specificity, and sensitivity:

$$Accuracy = (TP + TN)/N$$

$$Specificity = TP/N_1$$

$$Sensitivity = TN/N_0$$

Consider the binary case with an out-of-set sample size of N total data entries, containing N_0 observations from class 0 and N_1 observations of class 1. Define TP as the count of true positives or cases that are accurately predicted to be 1 when the known class is 1. TN is the count of true negatives or cases accurately predicted to be 0, when the known class is 0. These three classification metrics can be used to compare different algorithms. In cases where comparable accuracies are returned, sensitivity or specificity can be used to gain information on how often each class is accurately predicted.

2.2 Ensemble Learning

In supervised learning it is often observed that multiple models will be trained. Historically, once all models were evaluated, a common choice for the top performing model was that which maximizes out-of-set classification accuracy. This leads to the choice of a single model, which may produce unsatisfactory generalization to out-of-set

results. Modern learning algorithms incorporate ensemble learning to boost generalization performance. With antecedents in artificial intelligence, the basic principle of ensemble learning is that by using a weighted combination of classifiers, improved generalization to out-of-set results can be achieved (Freund and Schapire 1995).

(Dietterich 2002) provides a clear explanation of why ensemble models improve classification performance. Although, in general, ensembles are not driven by theoretical implications, by combining the output of learned models three types of problems can be overcome: statistical, computational, and representation. Statistical problems arise from searching wide argument spaces for specific features. It is possible to train several models that each lead to similar classification accuracy. Historically, a single model is chosen from the set of possible outcomes, risking that the chosen classifier may not perform well on out-of-set data. Instead, the statistical problem can be improved by using a combination of the top performing models, reducing the risk that the chosen model will not generalize well.

The computational problem relates to the possibility that training may produce classifiers that depend strongly on local features. Depending on how the model is constructed, particular features may be weighted strongly. If such features do not perform well the out-of-set data will not generalize well; hence, combinations of classifiers that differentially weight local features can improve generalization. This can be thought of as combining many approximations that find local minima to a problem, where the true global minimum is computationally infeasible to achieve.

The representation problem is the idea that the feature space available for model building is incapable of accurately representing the true underlying classes. In such cases using a weighted sum of classifiers expands the possible approximation space that can be achieved. The ensemble may then be capable of achieving closer approximations to the true generating mechanisms, achieving improved classification performance for a problem which is ill-suited to the available features.

The statistical, computational, and representation problems can each inflate the variance of the predictions while introducing bias in the learning procedure. Thus, ensemble methodologies are capable of improving predictive accuracy through variance reduction. The algorithm developed in Chapter 6 addresses each of the aforementioned problems. Empirical results are then used to demonstrate that ensembles both improve classification while simultaneously reducing the variance, and thus our uncertainty, of the predictions.

In addition to improving out-of-set predictions based on a single feature set, ensembles can be used to combine multivariate classifiers. Many common applications of supervised learning apply only to univariate cases, where data incorporated into the learning algorithm must be compatible, or produced from parallel learning algorithms. These are the types of applications that are widely available in the literature (Liu and Yao 1999; Rosen 1996). However, ensemble methodologies also allow for the extension to multivariate cases, where models built using different feature spaces can be combined to produce a final ensemble with improved generalization properties. This work attempts to exemplify such novel ensemble concepts that have yet to be investigated.

2.3 Functional Data Analysis

FDA is a growing statistical field that views data as functional representations rather than discrete observations (Ramsay 2006; Ramsay and Silverman 2007). For a set of n distinct observations $(y_1, y_2, \dots, y_n)$, we consider a latent function X that can be used to represent the set of observations by

$$y_j = X(t_j) + \varepsilon_j.$$

The observations are realizations of the function X at a given argument value t_j. This allows us to represent responses as a function $X(t)$ rather than as a set of discrete observations.

Considerations must be taken as to the best representation of the latent function; this includes the choice of smoothing parameters and basis functions. Let φ_k be the set of basis functions. The function X can be expressed as

$$X(t) = \sum_{k=1}^{K} c_k \varphi_k(t)$$

a linear expansion of K known basis functions. By representing the latent function through a basis expansion, the infinite-dimensional functional space is reduced to a finite-dimensional framework, where the dimension of the expansion is K. Determination of the dimension of the basis expansion relates to the degree of smoothness imposed when creating the functional approximations of each set of observations. A trade-off must be considered when choosing K such that the features present in the discretized observations

are retained while smoothing of the functions that allows for accurate representations of functional derivatives.

In addition to the choice of the dimension of the basis expansion, the choice of basis functions is also critical to obtaining accurate functional representations. Although bases such as polynomials or Fourier series have common applications, such bases may return inadequate derivative approximations. There is no single basis set that works well for all problems, and in general choice of basis functions and dimensionality are governed by derivative orders important to the analysis at hand.

One advantage of using FDA is that functional derivatives are easily computed once functional representations of the primary curves have been produced. Derivatives are quickly estimated using

$$D^{(j)}X(t) = \sum_{k=1}^{K} c_k D^{(j)} \varphi_k(t)$$

where $D^{(j)}$ represents the jth derivative order. Adequately produced functional representations to the raw curves will allow for derivatives to be incorporated into the data analysis.

Although derivative approximations can be produced from discretized measurements, proper choice of basis dimensionality and functions can result in smoothed derivative approximations unobtainable from conventional methods. This is why it is crucial to properly choose how the functional representation is constructed. Basis functions that give accurate approximations to the primary data may produce high-

frequency oscillations in derivatives. Oscillations in derivative approximations can have negative consequences on analysis, thus smoothing penalties and basis dimensionality are typically chosen based on which derivative orders are to be analyzed.

FDA provides a wide array of techniques once functional approximations have been produced. Many commonly applied statistical methodologies now have FDA counterparts that can produce improved results in comparison to discretized methods. Of interest to this work is the use of functional GLM and GAM (Ramsay et al. 2009), functional principal component analysis (FPCA) (Górecki and Krzyśko 2012), and functional supervised classification routines, primarily FuNC (Ferraty and Vieu 2006). Chapter 3 provides FDA of the SLE thermograms including decomposition by FPCA. Functional classification and ensemble strategies are evaluated in Chapter 5. A novel curve classification algorithm is then developed in Chapter 6 using FuNC as the primary methodology.

2.4 Plasma Thermograms

This dissertation presents and evaluates a novel classification problem that is based on plasma thermograms. Developed by (Garbett et al. 2007a; Garbett et al. 2007b), plasma thermograms provide a unique examination of the entire human plasma proteome through the use of differential scanning calorimetry (DSC). DSC is a thermoanalytical methodology that measures differential heat capacity changes between a sample and reference solution. Historically, DSC has been used to evaluate the thermodynamic stability of sample solutions, primarily interrogation of a single protein (Sturtevant 1987).

Modern applications include the thermodynamic characterization of novel therapeutic compounds, where DSC can be used to evaluate binding affinities and changes to the thermal stability of biomolecules (Höhne et al. 2013). The dissertation author has invested interest in calorimetric techniques, having used DSC to evaluate higher-order nucleic acid systems (Cashman et al. 2008; Chaires et al. 2014; Dettler et al. 2010; Dettler et al. 2011) and aid in the characterization of novel therapeutics (Freyer et al. 2007; Nagesh et al. 2010).

By measuring the differential energy input necessary to denature the sample in comparison with a reference solution, excess heat capacity curves known as thermograms are produced. Biochemically, thermograms can be deconvoluted to produce transition enthalpies and melting temperatures. Estimation of these two quantities based on thermodynamic models can be used to parse the free energy contributions that drive the thermal conformation change. For a single protein or nucleic acid system, this provides the Gibb's free energy, enthalpy, and entropy of the biomolecular conformational change from a native state (folded) to a denatured state (unfolded), and can be used to identify intermediate states if present.

Plasma thermograms are a unique and novel approach to the interrogation of the human plasma proteome. Plasma thermograms refer to the thermodynamic signature produced by subjecting a human plasma sample to thermal denaturation using DSC (Garbett et al. 2008). The resulting thermograms have been shown to correlate with disease states (Garbett et al. 2009) and are a promising methodology for improved diagnosis of diabetes (Garbett et al. 2013), several forms of cancer (Garbett et al. 2014;

Todinova et al. 2012; Zapf et al. 2011), and autoimmune disorders (Garbett et al. 2009). One major necessity to the development of plasma thermograms as a diagnostic measure is the statistical classification of a patient's health state based on the resulting thermogram signature. Such is the focus of several recent publications (Fish et al. 2010; Garbett and Brock 2016; Garbett et al. 2017; Garbett et al. 2015; Rai et al. 2013).

To date, such investigations have provided sub-par classification performance for identification of certain disease states, specifically the classification of autoimmune disorders. The primary data set of interest in this work is based on the classification of SLE against non-SLE patients (Garbett et al. 2008). The most recent statistical investigations of SLE plasma thermograms have provided classification accuracies no higher than 89%, which required combined statistical analysis of plasma thermograms in concert with traditional antibody testing (Garbett et al. 2017). There has been no investigation of plasma thermograms through FDA approaches, nor have derivative signatures and their potential improvements to classification been studied.

Chapter 4 will provide a modern statistical analysis of the SLE plasma thermogram dataset, reproducing much of the current literature results for reference in later chapters. Chapter 4 uses FDA to provide derivative signature approximations, with analysis based on the discretized sampling from functional representations. Chapter 5 presents FDA classification and ensemble strategies, demonstrating that SLE identification can be improved by the use of nonparametric and functional nonparametric classifiers. The continuous data collection provided by DSC is naturally extended to FDA, and through FPCA derivative curves will be shown to provide unique information

to the classification of disease states that has yet to be exploited in the literature. The algorithm developed in Chapter 6 will extend the use of FDA for the analysis of SLE plasma thermograms, demonstrating that multivariate analysis using derivatives can be used to boost classification accuracies beyond currently available methodologies. Classifiers based on FPC scores derived from FPCA of original, first, and second derivative curves will be studied in Chapter 7. Developing learning algorithms capable of evaluating the number of FPCs considered during the estimation of the classifiers results in improved classification performance over using traditional predictors.

Chapter 3

SLE PLASMA THERMOGRAM FUNCTIONAL DATA ANALYSIS

3.1 Systemic Lupus Erythematosus

SLE, commonly referred to as lupus, is an autoimmune disorder where the immune system targets healthy tissue throughout the body. Individuals with SLE may present symptoms including but not limited to joint pain, fever, rash, and ulcers. Symptoms often occur in cycles with periods of intense presentation and periods of remission. The cause of SLE is currently unknown, with both genetic and environmental factors believed to induce disease presentation. There is no cure for SLE but tailored treatment can produce effective control of symptoms, which improves quality of life (Hochberg 1997).

SLE treatment depends on accurate and timely diagnosis. SLE is commonly misdiagnosed because its symptoms relate closely to other autoimmune disorders such as fibromyalgia, Crohn's disease, psoriasis, and arthritis. Studies from the Lupus Foundation of America have reported misdiagnosis rates as high as 41%. This emphasizes a need for both improved primary care awareness and development of new assays to aid identification of SLE (Daly et al. 2017). Current diagnostic methods include a litany of tests based on blood and urine analysis, biopsies, and antinuclear

antibody testing (Hochberg 1997). Although these tests are useful, they can be invasive and time consuming, with accurate diagnosis taking months to years.

Plasma thermograms have been proposed as a new diagnostic tool for the identification of SLE (Garbett et al. 2015). The technique is minimally invasive, requiring only a blood sample, and can be performed in a short time frame. Plasma thermograms could provide an effective and quick assay for SLE identification. Used alongside current clinical techniques, thermograms have the potential to improve both the time frame and accuracy of SLE diagnosis.

The SLE plasma thermogram data set has been made available from Garbett N.C. at the University of Louisville. The data set provides 589 duplicated DSC thermogram signatures from 291 non-SLE controls and 298 SLE samples. Each scan contains excess heat capacity readings at 451 temperature points ranging from 45 – 90 ^{o}C. SLE and non-SLE patients may suffer from morbidities distinct from SLE. Thus, the non-SLE subgroup cannot be defined solely as healthy controls but as patients suffering from non-SLE illnesses.

3.2 The Classification Problem

For plasma thermograms to be employed effectively in the diagnosis of SLE, it is necessary to couple them with accurate classifiers. Primary interest is in production of a binary classification model capable of capturing SLE patients against non-SLE controls. Figure 1 depicts the raw plasma thermogram observations. The figures represent

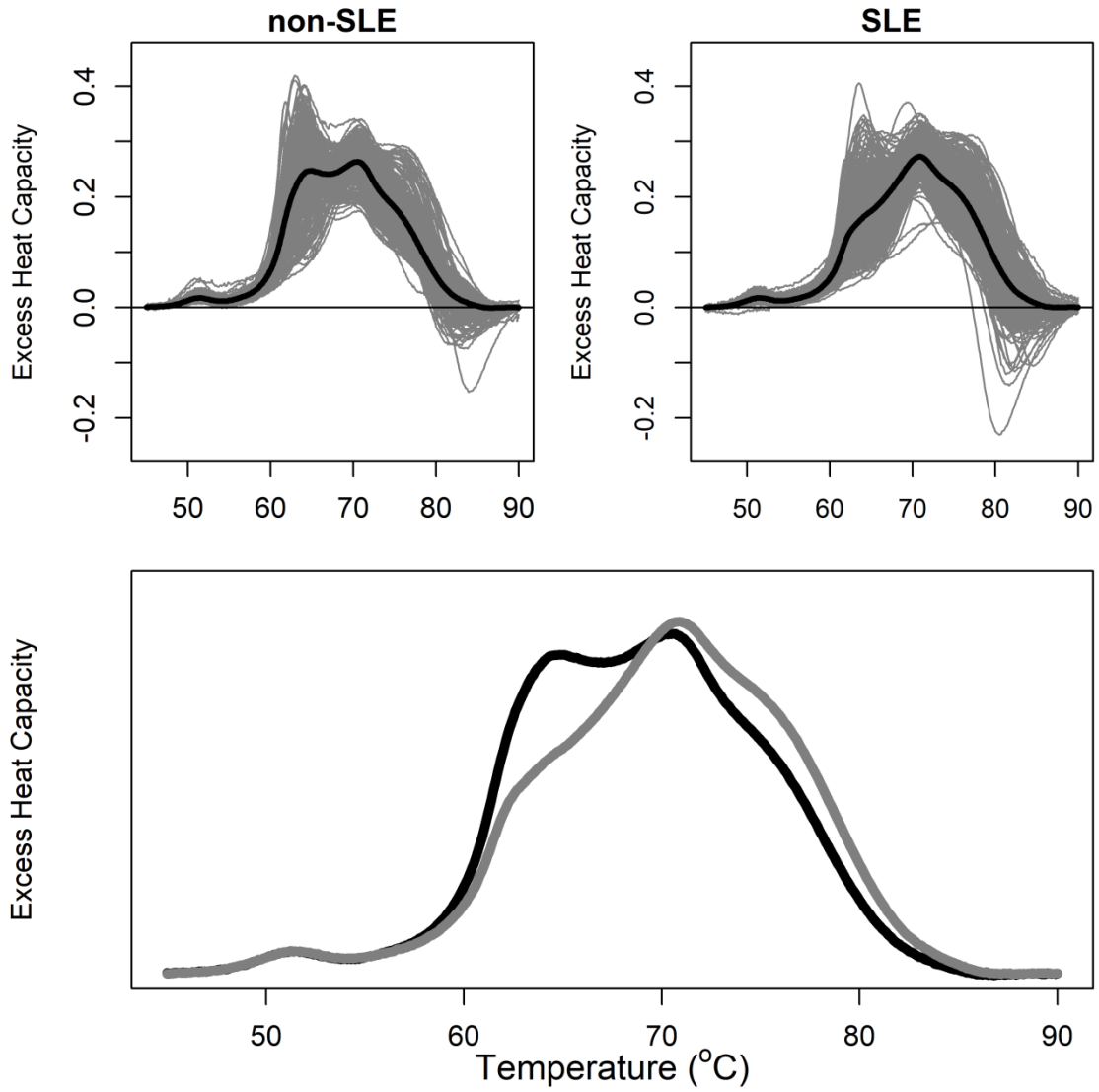

Figure 1. Raw observations from the SLE plasma thermogram data set. The top panels show raw curves given in grey with the mean of all observations given in black for non-SLE and SLE subsets. A comparison of mean signatures is given in the bottom panel: the non-SLE (black) signature differs from the SLE (grey) signature in two distinct temperature regions.

normalized DSC results, where the presented curve for each patient is the average of duplicated DSC experiments.

A comparison of the non-SLE and SLE signatures demonstrates a similarity in the distribution of maximum peak heights with both cases showing troughs at high temperature regions. An overlay of the mean curves from each class highlights the variability between SLE and non-SLE. The non-SLE plasma thermograms have a tendency to produce higher peak heights in the temperature region 60 – 70 ^{o}C. SLE plasma thermograms are shifted toward higher temperatures with maximum peak heights on average occurring between 70 – 80 ^{o}C.

The classification problem here aims to achieve high predictive accuracy of SLE cases against non-SLE alternatives. This problem has been investigated in previous research works (Fish et al. 2010; Garbett et al. 2015; Garbett et al. 2009; Garbett et al. 2008). Various classification algorithms have been utilized with subpar accuracies. Chapter 4 will present an in-depth look at contemporary classification algorithms and investigate the potential of using derivative-based predictors. This chapter presents the first use of FDA for representation of the plasma thermograms and the corresponding derivative curves.

3.3 Functional Representations

FDA will be used to produce data objects containing a set of random functions, X, paired with a binary response variable, Y: this will be termed a functional data object

(FDO). FDOs contain the information necessary to perform supervised classification algorithms. A preliminary step in constructing the FDO is optimizing the functional representations (X) of the raw data observations with respect to basis size and roughness penalties. For that purpose FDA functions from the R packages **fda** (Ramsay et al. 2014) and **fda.usc** (Febrero-Bande and de la Fuente 2012) were used.

A generalized cross validation (GCV) procedure is available through **fda.usc** that allows for simultaneous evaluation of the basis size and roughness penalties. Basis size refers to the number of basis functions (K) to be used in the linear expansion of the latent function. The roughness penalty (λ) is a smoothing parameter used to penalize the curvature of derivative curves. A penalty term based on the integrated squared mth-derivative of the function is given by

$$PEN_m = \int \left[D^{(m)}X(s)\right]^2 ds.$$

(Ramsay and Silverman 2005) define the penalized residual sum of squares

$$PENSSE_\lambda(X|Y) = [Y - X(t)]'W[Y - X(t)] + \lambda * PEN_2(X)$$

where W is a weight matrix. $PENSSE$ allows one to estimate the function X over the space for which the penalty term is defined. Typically curvature is penalized corresponding to $m = 2$. $PENSSE$ can be used on higher derivative orders if one wishes to include the analysis of a particular derivative in their analysis.

Smoothing parameters control the trade-off between closeness of the data to the functional representation and data averaging, which provides smooth derivative

information. When $\lambda \to 0$, the curvature of X approaches an interpolant of the raw curves, providing exact approximations at the observation points but high variability in derivative approximations. Alternatively, as $\lambda \to \infty$ the functional representation will converge to the standard linear regression through the observations where $PEN_2(X) = 0$. This provides biased estimates at each observation point but reduces the variability of derivative estimates.

The GCV routines stem from smoothing B-splines (Golub et al. 1979), and construct a generalized metric for evaluating the goodness of functional representations. GCV within **fda.usc** is evaluated over a grid of both basis sizes (K) and roughness penalties (λ). The goal is to produce functional representations of the SLE plasma thermograms that accurately approximate the observed thermogram points, while allowing for smoothed evaluation of derivatives. Figure 2 shows functional approximations of the SLE plasma thermograms. The functions were produced using B-splines that were unsmoothed ($K = 451, \lambda = 0$) and GCV optimized ($K = 150, \lambda = 0$). Classifiers will be constructed using both unsmoothed and GCV optimized functional representations for comparison. Fourier basis functions were also considered but produced higher optimized GCV. Smoothing penalties based on derivatives will not be presented as they did not improve GCV.

Figure 2 presents the unsmoothed and GCV functions side by side for the original, first, and second derivative curves. $X(t)$ stands for the excess heat capacity at temperature t, with $X'(t)$ and $X''(t)$ corresponding to the first and second derivatives with respect to temperature. No significant changes at this scale can be observed in the

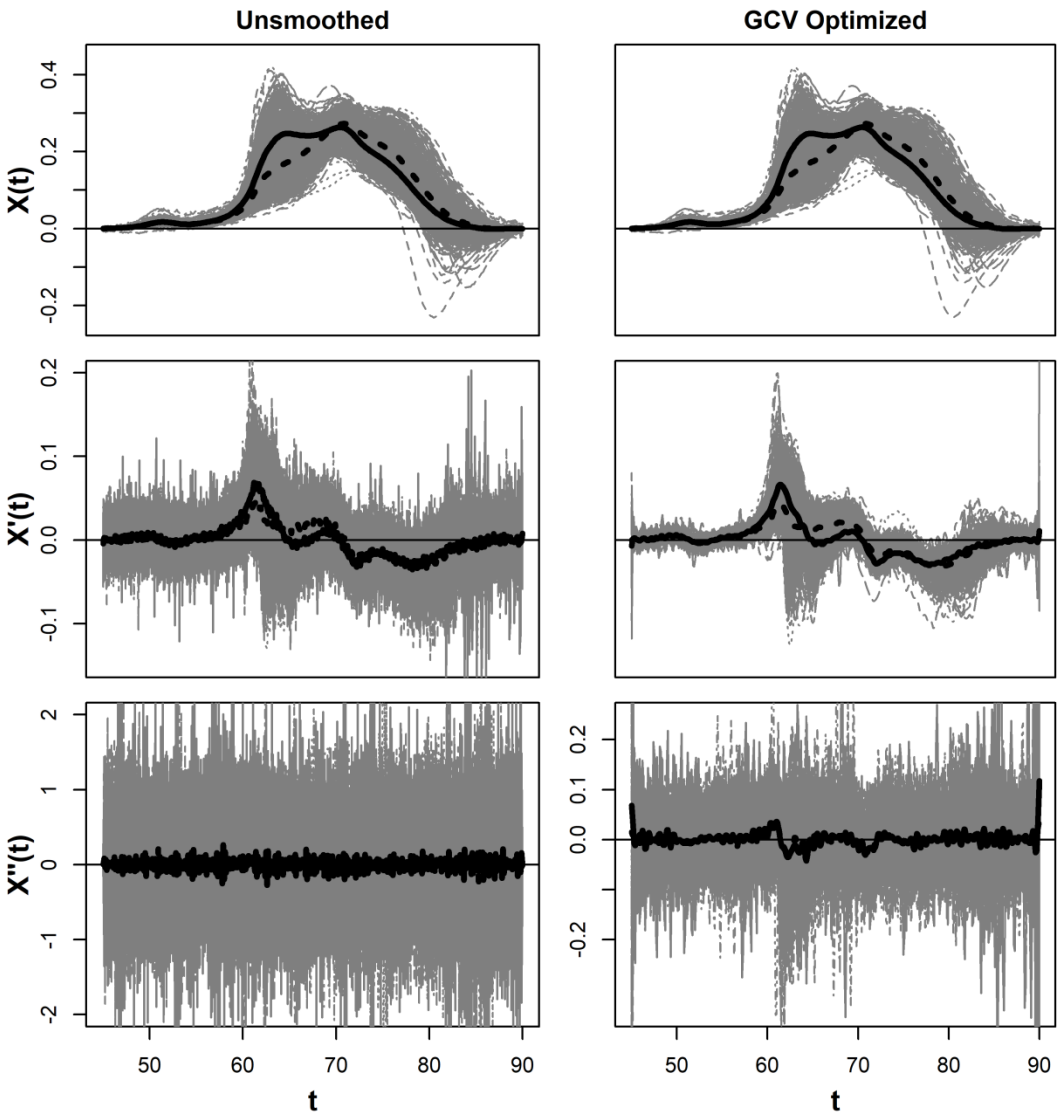

Figure 2. Comparison of unsmoothed and GCV optimized functional data representations of SLE plasma thermograms. Shown are the original (top), first (middle), and second derivative (bottom) curves. SLE (solid black) and non-SLE (dashed black) mean curves are provided. Unsmoothed refers to B-splines created with a basis expansion of size $K = 451$ and roughness penalty of $\lambda = 0$. GCV resulted in an optimized basis expansion of size $K = 225$ and roughness penalty of $\lambda = 0$. The influence of the basis expansion size and roughness penalty on the approximation of derivative curves can be clearly observed.

original curves. However, an investigation of the derivatives clearly shows the impact of the GCV optimization. First derivatives from unsmoothed approximations show large oscillations; the oscillations are significantly reduced in the GCV representation, with small numerical noise appearing near the endpoints. Second derivatives from unsmoothed representations are scattered with no distinguishable mean pattern. Smoothed functions show improvements in the interpretability of second derivatives with distinct peaks appearing near critical regions of the thermograms (60 – 80 ^{o}C).

FDA has provided the construction of functional random variables that can be paired with their corresponding class identifiers for the creation of FDOs. The FDOs corresponding to the original curves from unsmoothed and GCV optimized functional approximations will be denoted as SLE FDO and GCV FDO, respectively, from here in. These primary FDOs can be used to produce derivative FDOs, as visualized in Figure 2. This expands the potential set of information for classification to original data observations and their derivative approximations. Figure 3 illustrates the first derivative approximations generated from the GCV FDO. The curves are partitioned into SLE and non-SLE cases. A comparison of the curves from each class shows clear distinctions in the mean signal over a range of temperatures from 55 – 85 ^{o}C. A significant difference in the mean signal occurs between 60 – 70 ^{o}C. Near 65 ^{o}C, a change occurs in the class having the higher mean signal. The first derivatives show only small oscillations near the temperature range endpoints.

Figure 4 presents the second derivative approximations; it is evident that numerical noise is present even after GCV optimization of the functional approximations.

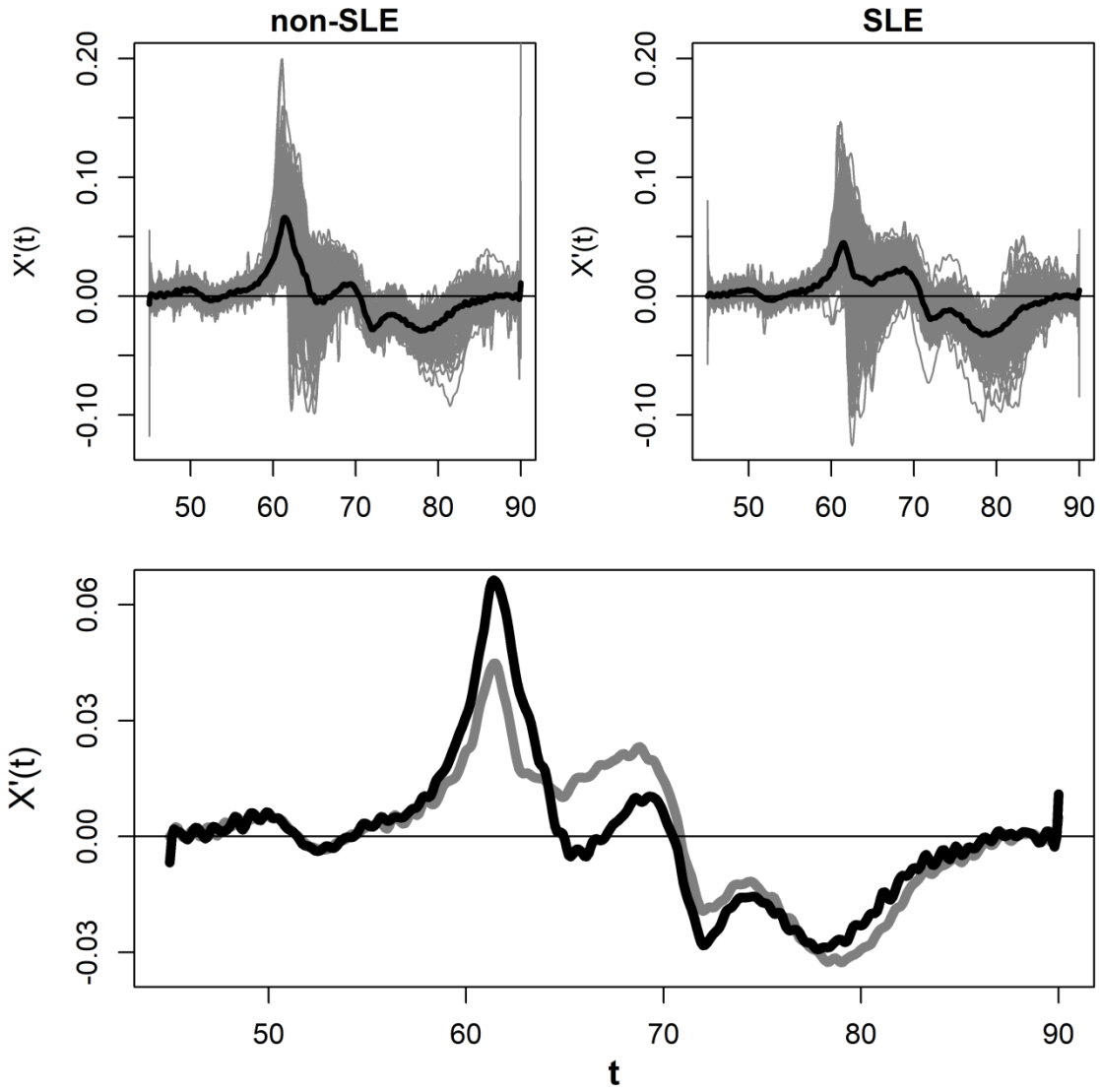

Figure 3. First derivative approximations of the SLE plasma thermogram data set. Plots were generated from functional approximations using the GCV FDO. The top panels show first derivative curves given in grey with the mean of all observations given in black for non-SLE and SLE subsets. A comparison of mean first derivative signatures is given in the bottom panel: the non-SLE (black) signature differs from the SLE (grey) signature in the region 60 – 70 ^{o}C. A change in which class has the higher peak height occurs near 65 ^{o}C.

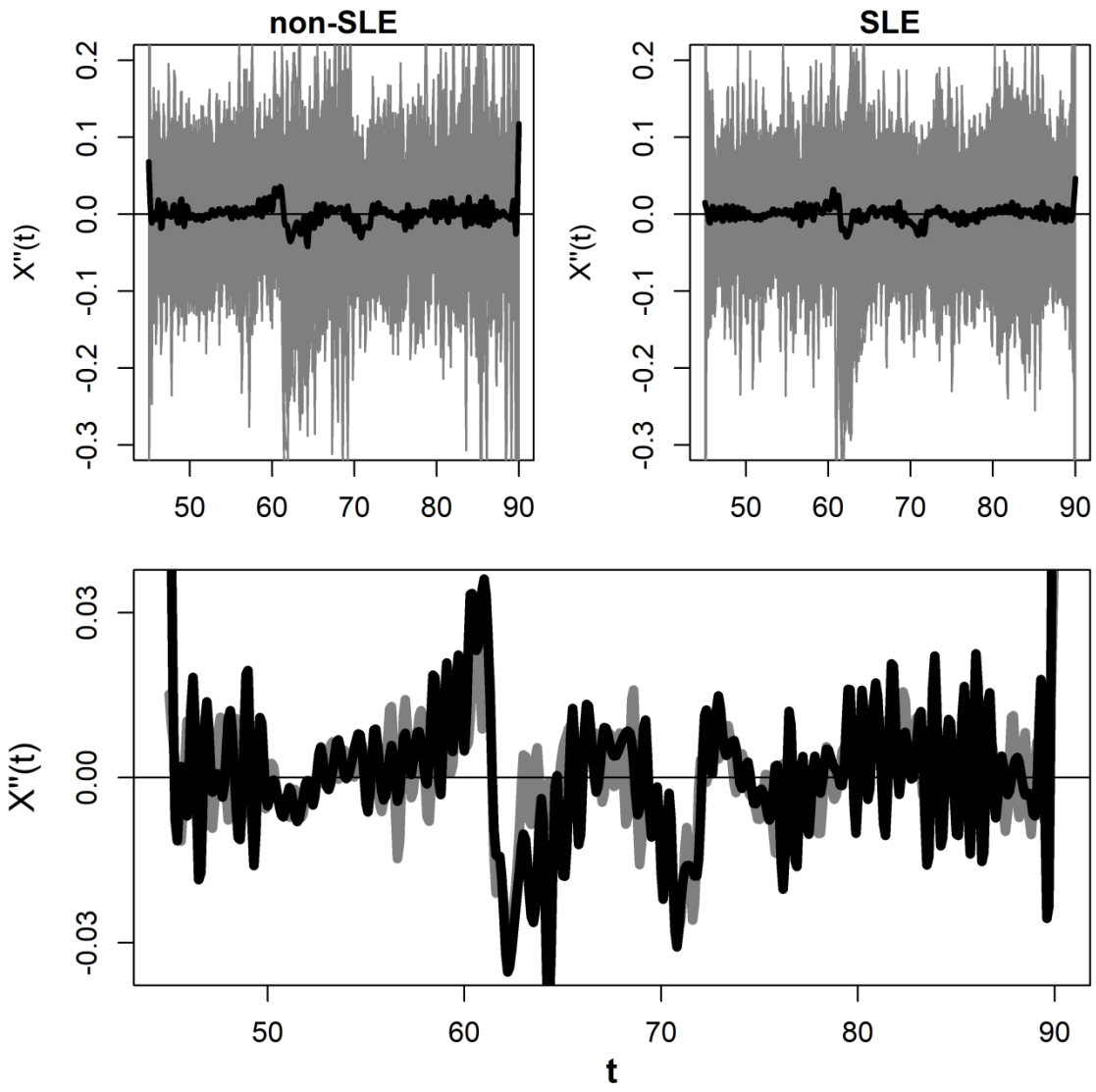

Figure 4. Second derivative approximations of the SLE plasma thermogram data set. Plots were generated from functional approximations using the GCV FDO. The top panels show second derivative curves given in grey with the mean of all observations given in black for non-SLE and SLE subsets. A comparison of mean second derivative signatures is given in the bottom panel: the mean non-SLE (black) curve has a clear pattern of distinction from the mean SLE (grey) curve near 65 ^{o}C. Noise is reduced but still clearly present in the approximation of the second derivative curves.

Although noisy, the derivative approximations still offer insight into the differences between SLE and non-SLE. Specifically, it is clear that there are signal differences between SLE and non-SLE mean curves: a difference in the mean curves appears near 65 ^{o}C, matching the temperature at which the maximum peak height of the class switches.

This analysis highlights important details on the inclusion of plasma thermogram derivatives in the classification of SLE. It is clear that FDA is a tractable and enticing tool for improving the investigation of plasma thermograms. FDA results in excellent functional representations with clear improvements to the visual inspection of derivative information. Exploratory analysis of the derivative curves suggests they can be used to gain additional signal for classifying SLE against non-SLE alternatives. Unique to this dissertation will be the investigation of contemporary learning methodologies based on plasma thermogram derivative curves.

3.4 Functional Principal Component Analysis

PCA is an integral step in data analysis as it allows one to explore the features characterizing the variations inherent to a data set, while quantifying the covariance structure. PCA analysis compliments the typical evaluation of variance-covariance of predictors and allows for reduction to the feature dimensions providing the greatest explanation of variance within the data set (James et al. 2013; Ramsay and Silverman 2005). The typical multivariate standpoint considers n centered response values $(y_1, y_2, \dots, y_n)$ produced from p-dimensional feature vectors $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})'$. A

central statistical concept is to exploit a linear combination of the features using weighting coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_p)'$ such that for $i = 1, \ldots, n$

$$y_i = \sum_{j=1}^{\mathrm{p}} \beta_j x_{ij} = \boldsymbol{\beta}' \boldsymbol{x_i}.$$

PCA determines a set of normalized weight vectors $\boldsymbol{\xi_j} = (\xi_{1j}, \xi_{2j}, \ldots, \xi_{pj})'$ that maximize the variation in the response. This is performed by evaluating principal component scores

$$y_{i1} = \sum_{j=1}^{\mathrm{p}} \xi_{j1} x_{ij} = \boldsymbol{\xi_1}' \boldsymbol{x_i}$$

such that the mean square principal component $(MSPC)$ score

$$\mathrm{MSPC}_1 = \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} y_{i1}^2$$

is maximized, under the constraint

$$\sum_{j=1}^{\mathrm{p}} \xi_{j1}^2 = \|\boldsymbol{\xi_1}\|^2 = 1.$$

This identifies the strongest mode of variation within the feature set. Additional weight vectors $\boldsymbol{\xi_m}$ for $m = 2, \ldots, p$ are found analogously to the above with the additional $m - 1$ constraints for $k < m$

$$\sum_{j=1}^{p} \xi_{jk}\xi_{jm} = \boldsymbol{\xi_k}'\boldsymbol{\xi_m} = 0.$$

This produces a set of orthonormal weighting vectors, each capturing the next most important mode of variation within the feature set. Importantly, each vector will be uncorrelated with all others, providing a methodology for producing uncorrelated feature sets. Under standard multivariate analysis, PCA is produced from an eigen-decomposition of the variance-covariance matrix. This results in a set of eigenvectors $\boldsymbol{\xi_j}$ and corresponding eigenvalues ($\mathrm{MSPC_j}$).

However, under the context of FDA, we no longer consider the observed features $\boldsymbol{x}$ as discretized but as realizations of the function $X(t)$. From the FDA standpoint, the summation of a linear combination of weighted coefficients with discretized features becomes the evaluation of an integral. For general linear models, the product of a weighting function, $\beta(t)$, is taken with the functional representations, $X(t)$. The responses can now be written as

$$y_i = \int \beta(s)X_i(s)\, ds$$

where $X_i(t)$ represents the latent functional representation of the ith response.

For FPCA, the weighting vectors become weighting functions, $\xi_j(t)$. Decomposition into the maximized variance components is done analogously to the above discussion with summation replaced with integration. The mth functional principal scores

$$y_{im} = \int \xi_m(s) X_i(s)\, ds$$

are used to maximize the mth functional mean square principal component

$$\text{fMSPC}_\text{m} = \frac{1}{\text{n}} \sum_{i=1}^{\text{n}} y_{im}^2$$

subject to the constraint now defined under the continuous norm

$$\int \xi_m^2(s)\, ds = \|\xi_m\|^2 = 1.$$

For all weight functions beyond the first, there are an additional $m - 1$ constraints to ensure orthogonality of the functions, so that

$$\int \xi_k(s) \xi_m(s)\ ds = 0$$

when $k < m$.

The material presented above provides an analogous setup of PCA under a functional context. FPCA produces a set of eigenfunctions, $\xi_j(t)$, along with corresponding eigenvalues, fMSPC_j. Unlike traditional eigen-decomposition used for PCA, solutions are found based on Karhunen-Loeve transformations (Dony 2001). The resulting FPCA allows one to investigate the sources of variation based on eigenfunctions, producing a smoothed interpretation of the sources of variation within the curves. Additionally, sources of variation within derivative approximations can also be studied, allowing for one to evaluate variations in the rate of changes as well.

FPCA was applied to the SLE plasma thermogram set to produce principal component decompositions of the primary FDOs along with their first and second derivative approximations. Figure 5 depicts the FPCA decomposition for the first four eigenfunctions of the primary SLE functional approximations. The figure presents eigenfunctions (components) generated using the entire SLE plasma thermogram data set and partitions into SLE and non-SLE cases. The first component explains 63.5% of the variation observed from all SLE plasma thermograms. The trough near 65 ^{o}C is lower for SLE than non-SLE curves, confirming more variability in the peak densities in this region for non-SLE patients. Similarly, the variation in SLE-patients is higher near 75 ^{o}C. This region corresponds to where many SLE curves are maximized, with higher peak heights on average then non-SLE curves.

The second component shows remarkable differences between SLE and non-SLE cases: it explains 14.8% of the variance observed from all SLE plasma thermograms. The region from 70 – 80 ^{o}C displays a stark difference in the explanation of variance between non-SLE and SLE cases. Within this region, the non-SLE cases have eigenfunction amplitude near 2, while the SLE cases show a significant dip in variation with amplitude near 0.5. This region corresponds to the maximum peak height of the SLE curves, and suggests that SLE patients have significantly reduced variation within this temperature region. The third and fourth components comprise 10.9% and 5.7% of the total observed variance. The third component shows oscillations in the variation of the curves over a wide temperature range of 60 – 85 ^{o}C. The fourth component picks up denser oscillations in the variation over the same temperature regions.

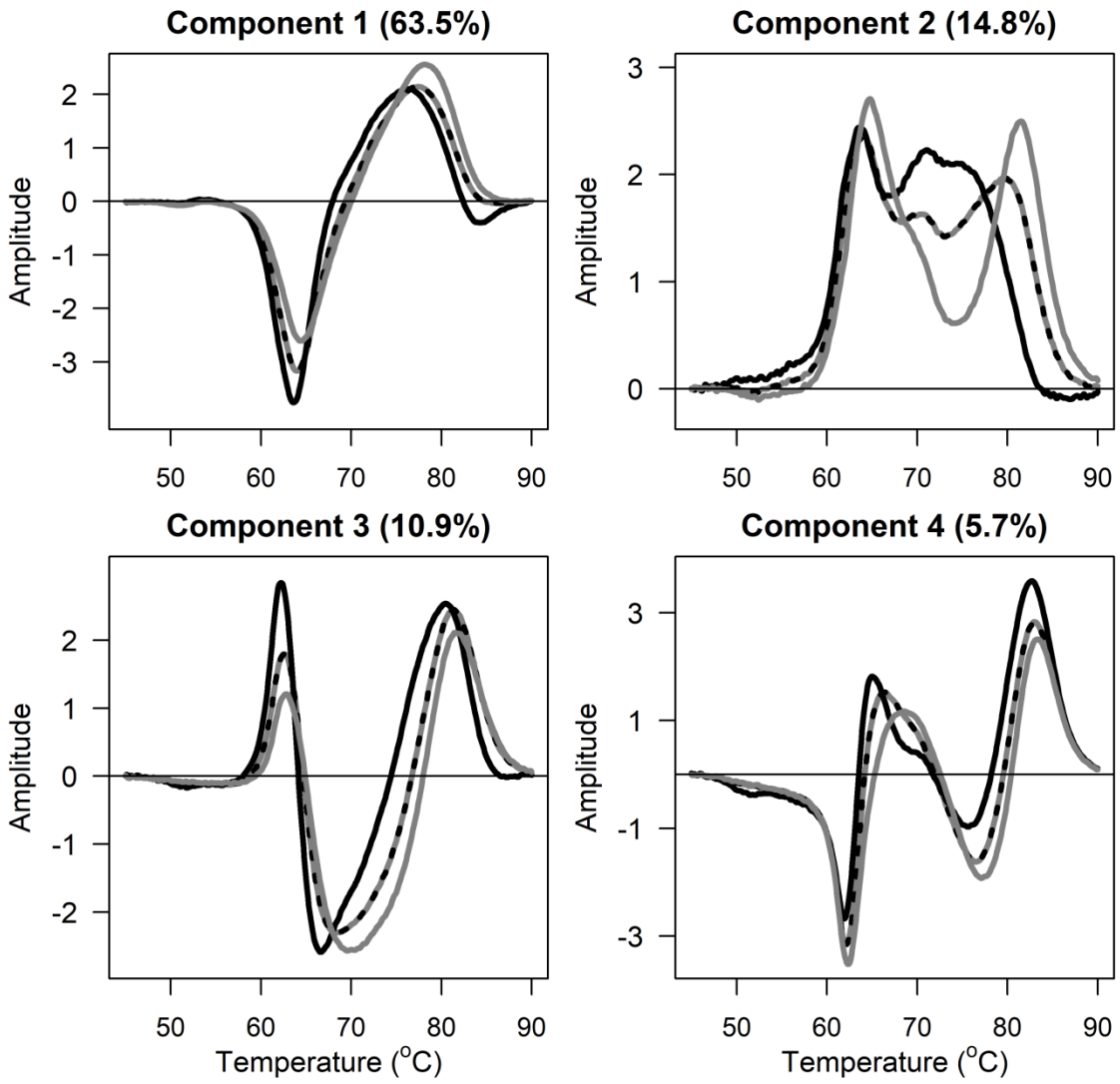

Figure 5. FPCA of SLE plasma thermogram original curves. The first four eigenfunctions are shown along for non-SLE (black solid), SLE (grey solid), and pooled curves (black and gray dash). The presented percentages correspond to the proportion of variance explained by each component.

Figure 6 presents the first two eigenfunctions resulting from FPCA of the first and second derivative approximations of the SLE plasma thermograms. The third component of the primary curves mimics the rate of changes found by FPCA evaluation of the first derivative approximations; the same three unique regions of variability, separated as less than 65 ^{o}C, 65 – 75 ^{o}C, and greater than 75 ^{o}C, around found in both. The first component of the first derivative explains 41.0% of the variation within the derivative approximations, increasing from only 10.9% of the variation explained from the third component of the primary curves. This suggests that analysis of the derivatives curves may be more sensitive to these variations and could improve overall analysis of the SLE plasma thermograms.

The second eigenfunction of the first derivatives shows sharp oscillations near the temperature where non-SLE peaks reach a maximum. This component explains 16.5% of the variation within the first derivative approximations. The first component of the second derivative explains only 7.4% of the variation and seems strongly influenced by the numerical noise at the temperature endpoints. The second component explains 7.0% and has a similar spike in variation in the critical region of 60 – 70 ^{o}C. Although unlikely to aid greatly in the classification of SLE vs. non-SLE alternatives, FPCA of the second derivative curves makes it clear that components distinct from random noise are present within the observations.

This analysis represents the first deconvolution of SLE plasma thermograms using FDA methodologies. Recent investigation of the SLE plasma thermograms presented a PCA breakdown of the primary results (Garbett and Brock 2016). The analysis presented

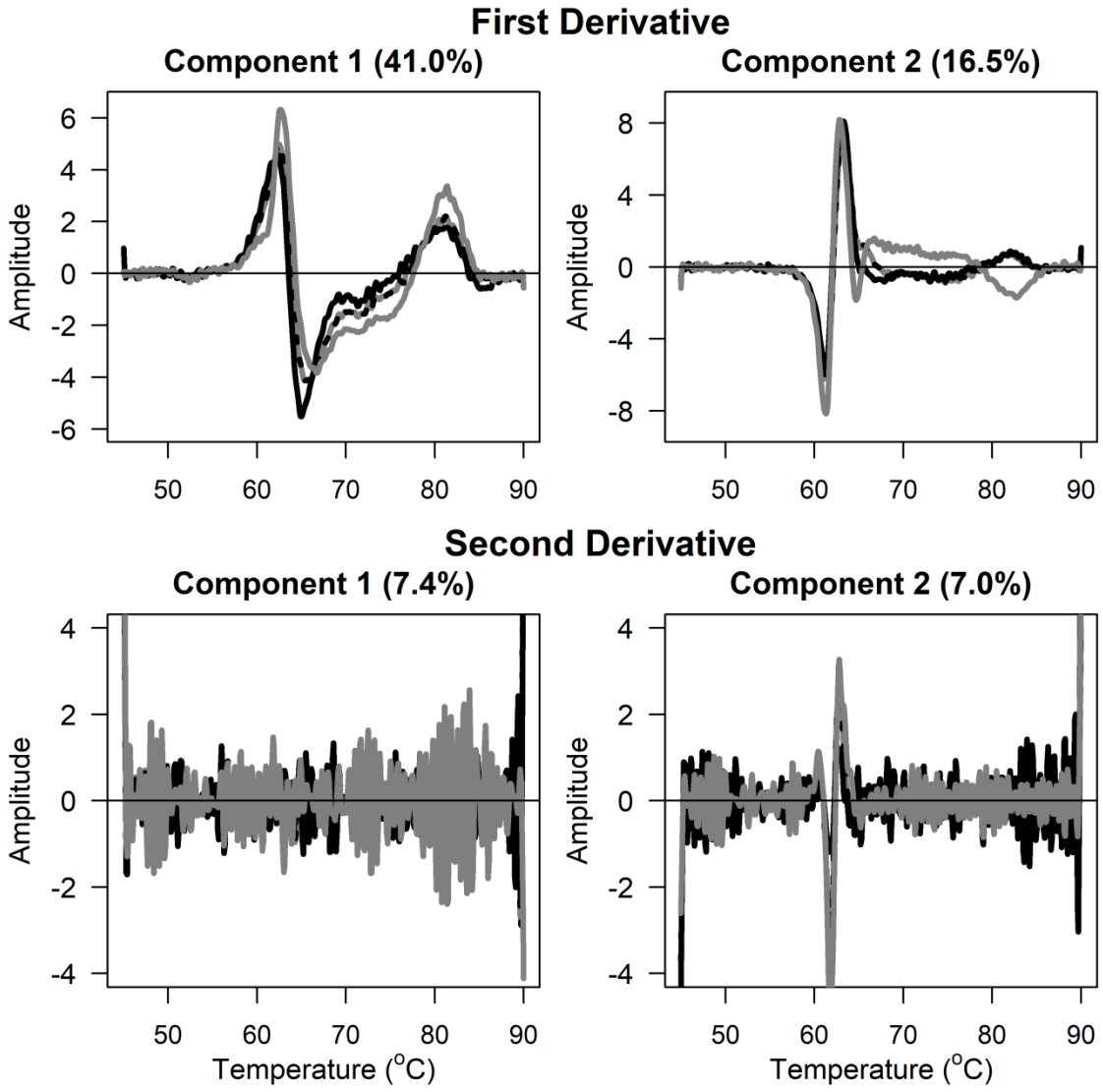

Figure 6. FPCA of first and second derivative approximations of SLE plasma thermograms. The first two eigenfunctions are shown for each derivative. Each graph shows the eigenfunction resulting from analysis of non-SLE (black solid), SLE (grey solid), and pooled curves (black and gray dash). The given percentages are the proportion of variance explained by the component.

here agrees with previous findings, but is capable of extending the analysis to derivative approximations. It is clear from FPCA that derivatives contain distinct patterns of variation distinct from that of the primary curves. This suggests that inclusion of the derivatives in supervised classification algorithms is likely to improve classification performance.

PCA based LR was studied in (Garbett and Brock 2016) resulting in a classification accuracy of only 70%. The authors used only the first 6 components during estimation of LR classifiers. FPCA based supervised learning algorithms will be investigated further in Chapter 7. Rather than using a predetermined number of FPCs, supervised learning will be evaluated over a grid of FPCs to produce high accuracy classifiers overlooked in recent literature.

Further directions involving FPCA would involve unsupervised learning methodologies such as clustering based on principal components (James et al. 2013). Instead, this thesis will focus on supervised methodologies. The functional representations of the SLE plasma thermogram will be used to conduct contemporary supervised classification based on discretization of the functional approximations and their derivatives. Functional supervised classification will then be presented, with a focus on NC.

Chapter 4

SUPERVISED CLASSIFICATION OF SLE PLASMA THERMOGRAMS

4.1 Introduction

This chapter evaluates contemporary supervised classification algorithms. Classification of the SLE plasma thermograms was performed using various LR estimators: maximum likelihood (ML), LASSO, RIDGE, and elastic-net (ENET) (James et al. 2013). The analysis includes an investigation and discussion of adaptive-LASSO (Zou 2006), adaptive-ENET (Zou and Zhang 2009), and least squares approximation (LSA) for adaptive LASSO (Wang and Leng 2007). Parametric models for classification of SLE vs. non-SLE alternatives use thermogram readings as predictors.

LDA and quadratic discriminant analysis (QDA) will also be considered; both models use Bayes' theorem to produce linear combinations of predictors (James et al. 2013). NC is explored using k-nearest neighbors (KNN) classifiers. Cross-validation for testing model performance will be done through KCV. A partitioning size of 10 is used for the folds, and to stabilize the variance, the KCV algorithm is repeated 20 times. To ensure proportionality of the sampling, stratified folds were created using built-in functions of the **caret** (Kuhn et al. 2015) package. KCV is summarized by the three classification metrics as defined in Chapter 2: accuracy, sensitivity, and specificity.

The analysis uses discretized observations from SLE FDO and GCV FDO along with their derivative approximations as predictors. Function discretization presents problems known as the curse of dimensionality. How to sample from the functional approximations to produce optimal solutions is a difficult problem discussed in the literature (Ferraty and Vieu 2003; Verleysen and François 2005). To evaluate how sampling from the functions affects model building, two sets are drawn from each FDO. The first set (FULL) uses the original temperature grid of 45 – 90 ^{o}C with sampling every 0.1 ^{o}C. A second set (TRUNC) reduces the dimensionality of the predictor set by sampling every 0.5 ^{o}C. These two sampling methods were used to construct four unique discretized predictor matrices from the SLE plasma thermograms: SLE FULL, GCV FULL, SLE TRUNC, and GCV TRUNC. Predictor matrices were generated for the original curves along with first and second derivative approximations.

This analysis represents the first in-depth investigation of SLE plasma thermogram and their derivatives using the aforementioned statistical techniques. The predictive performance of these contemporary methodologies will be presented and the importance of derivative approximations will be discussed. The results will illustrate the difficulties inherent to the SLE thermogram classification problem promoting the development of new predictive algorithms based on FDA and ensemble methodologies.

4.2 Logistic Regression Estimators

LR models were constructed for the classification of SLE vs. non-SLE alternatives. LR is formulated as a linear model for the log-odds (or logit):

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j .$$

LR is a generalized linear model (GLM) under the binomial distribution with a logit link function. The log-odds function produces a linear model in the predictor coefficients and the log-likelihood is expressed assuming responses are Bernoulli distributed random variables. ML estimation is employed to calculate optimized parameters based on Fisher scoring or Newton-Raphson methods (Czepiel 2002; Fisher 1925; Ratcliffe et al. 2002). The log-odds function provides estimates of the probability an individual belongs to class 1 given the information provided by X, i.e. $p(x) = P(Y = 1|X)$. In what follows, LR was performed using the standard **glm** function in R.

Table 1 depicts the classification performance using the SLE FULL discretized predictors. Supervised learning for nine classifications methods is presented along with the resulting accuracy, specificity, and sensitivity from KCV. The predictors X are realizations from FDA approximations $X(t)$; when the functions are discretized to the original (FULL) temperature grid, 451 predictors from each curve are produced. This causes a relatively high-dimensional state to the analysis. More importantly, high collinearity is present between the predictors causing variance inflation. LR produces 72.0% mean out-of-set test accuracies for predicting SLE from non-SLE alternatives using the original data points (SLE FULL). The resulting models have 75.0% sensitivity and 68.8% specificity indicating that SLE patients are captured at a higher rate than non-SLE cases.

Original Curves			
Method	Accuracy	Sensitivity	Specificity
LR	0.720 (0.060)	0.750 (0.085)	0.688 (0.092)
RIDGE	0.876 (0.043)	0.874 (0.062)	0.877 (0.062)
ENET	**0.905** (0.040)	**0.904** (0.057)	**0.905** (0.057)
adap-ENET	0.874 (0.046)	0.882 (0.059)	0.866 (0.070)
LASSO	0.892 (0.042)	0.895 (0.060)	0.890 (0.060)
adap-LASSO	0.864 (0.047)	0.870 (0.064)	0.858 (0.070)
LDA	0.740 (0.056)	0.756 (0.083)	0.724 (0.081)
QDA	DNC	DNC	DNC
KNN	0.762 (0.052)	0.722 (0.079)	0.804 (0.074)
First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.718 (0.055)	0.735 (0.083)	0.700 (0.086)
RIDGE	**0.916** (0.034)	0.915 (0.049)	**0.917** (0.054)
ENET	0.909 (0.036)	0.912 (0.054)	0.906 (0.053)
adap-ENET	0.892 (0.042)	0.897 (0.055)	0.887 (0.061)
LASSO	0.900 (0.038)	0.905 (0.052)	0.896 (0.058)
adap-LASSO	0.879 (0.046)	0.887 (0.057)	0.871 (0.070)
LDA	0.741 (0.056)	0.756 (0.083)	0.725 (0.083)
QDA	DNC	DNC	DNC
KNN	0.908 (0.039)	**0.945** (0.040)	0.870 (0.074)
Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.721 (0.058)	0.735 (0.086)	0.706 (0.083)
RIDGE	0.860 (0.045)	0.876 (0.057)	**0.844** (0.070)
ENET	0.854 (0.049)	0.869 (0.060)	0.838 (0.076)
adap-ENET	0.845 (0.050)	0.857 (0.063)	0.833 (0.078)
LASSO	0.853 (0.047)	0.866 (0.061)	0.839 (0.076)
adap-LASSO	0.841 (0.049)	0.851 (0.062)	0.831 (0.076)
LDA	0.796 (0.052)	0.809 (0.072)	0.783 (0.082)
QDA	DNC	DNC	DNC
KNN	**0.876** (0.039)	**0.933** (0.042)	0.817 (0.072)

Table 1. SLE FULL classification performance of the nine classifiers as given by the accuracy, specificity, and sensitivity. Mean and standard deviation for each metric is given for the original curves and their first and second derivative approximations. DNC indicates that a model failed to converge.

Resulting models from LR using predictors from first and second derivative approximations are also given in Table 1. LR has reduced predictive performance using derivative approximations, with the mean out-of-set accuracies reducing to 71.8% and 72.1% for the first and second derivatives, respectively. The effects of smoothing using B-spline basis reduction are assessed in Table 2, where the classification results using the GCV FULL discretization grid are presented. LR demonstrates a gain in predictive accuracy from the original curves using points sampled from smoothed functions, increasing to an out-of-set mean accuracy of 80.3%. The GCV FULL grid also shows improved classification performance to SLE FULL using the derivative approximations with mean test set accuracies of 79.0% and 79.1% for the first and second derivative predictors, respectively.

Penalized LR was investigated for improvements to classification performance. Penalized models estimate the coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_p)'$ that minimize the penalized likelihood specification (Tibshirani 1996; Tibshirani 2011). Different penalty terms produce variants of penalized estimation. Given below is the penalized likelihood problem for logistic regression including both l_1- and l_2-penalization terms.

$$\frac{1}{N}\sum_{i=1}^{\mathrm{N}}\left[\mathrm{y_i}\left(\beta_0 - \mathrm{x_i^T}\beta\right) - \log\left(1 + \mathrm{e}^{\left(\beta_0 - \mathrm{x_i^T}\beta\right)}\right)\right] + \ (1-\alpha) * \lambda_1 \sum_{\mathrm{j}} \beta_j^2 + \ \alpha * \lambda_2 \sum_{\mathrm{j}} \left|\beta_j\right|$$

This form of penalized minimization problems have been deeply investigated with computational solutions available in numerous R packages (Goeman 2010; Yang et al. 2017). This chapter is primarily based on **glmnet** (Friedman et al. 2009), which is well known for its computational speed based on the coordinate descent algorithm. This

Original Curves			
Method	Accuracy	Sensitivity	Specificity
LR	0.803 (0.053)	0.828 (0.090)	0.778 (0.117)
RIDGE	0.829 (0.050)	0.833 (0..070)	**0.873** (0.060)
ENET	**0.855** (0.047)	0.861 (0.062)	0.849 (0.069)
adap-ENET	0.850 (0.049)	0.853 (0.066)	0.847 (0.069)
LASSO	0.853 (0.047)	0.859 (0.063)	0.845 (0.070)
adap-LASSO	0.849 (0.048)	0.856 (0.063)	0.841 (0.072)
LDA	0.853 (0.042)	**0.867** (0.060)	0.838 (0.065)
QDA	DNC	DNC	DNC
KNN	0.762 (0.052)	0.722 (0.079)	0.802 (0.075)
First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.790 (0.053)	0.807 (0.076)	0.773 (0.086)
RIDGE	**0.877** (0.041)	**0.879** (0.056)	**0.873** (0.060)
ENET	0.875 (0.042)	**0.879** (0.055)	0.871 (0.065)
adap-ENET	0.869 (0.049)	0.876 (0.057)	0.862 (0.073)
LASSO	0.875 (0.041)	**0.879** (0.053)	0.870 (0.064)
adap-LASSO	0.869 (0.041)	0.878 (0.055)	0.861 (0.060)
LDA	0.854 (0.044)	0.865 (0.061)	0.842 (0.064)
QDA	DNC	DNC	DNC
KNN	0.774 (0.054)	0.773 (0.073)	0.774 (0.079)
Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.791 (0.052)	0.806 (0.074)	0.776 (0.082)
RIDGE	0.861 (0.043)	0.872 (0.056)	0.849 (0.066)
ENET	**0.866** (0.040)	0.878 (0.056)	0.853 (0.061)
adap-ENET	0.863 (0.041)	0.871 (0.061)	**0.855** (0.062)
LASSO	0.865 (0.040)	0.877 (0.056)	0.853 (0.061)
adap-LASSO	0.859 (0.042)	0.867 (0.062)	0.851 (0.061)
LDA	0.856 (0.044)	0.871 (0.062)	0.841 (0.065)
QDA	DNC	DNC	DNC
KNN	0.865 (0.045)	**0.947** (0.037)	0.781 (0.079)

Table 2. GCV FULL classification performance of the nine classifiers as given by the accuracy, specificity, and sensitivity. Mean and standard deviation for each metric is given for the original curves and their first and second derivative approximations. DNC indicates that a model failed to converge.

package does not allow for differential weighting of the penalty terms, enforcing that $\lambda_1 = \lambda_2$. Hence, α controls the differential weight of l_1- and l_2-penalty terms.

When $\alpha = 0$, this produces l_2-penalized LR commonly known as the RIDGE estimator (Hoerl and Kennard 1970). The l_2-penalization produces models whose coefficients are decreased in magnitude in a regularized fashion. Solutions from RIDGE have reduced coefficient magnitudes, but the technique does not strictly enforce coefficients to be zero. The LASSO estimator refers to l_1-penalized LR when $\alpha = 1$. LASSO enforces predictor selection, having the property that insignificant coefficients are reduced to exactly zero. This causes LASSO to produce sparse models, where insignificant parameter coefficients are set to zero, thus removing the corresponding predictor from the classifier. LASSO can perform both predictor selection and parameter estimation simultaneously, improving model performance and model interpretability (Tibshirani 1996). ENET uses a linear combination of l_1 and l_2-penalization to produce models that are both sparse in parameters with regularized coefficients from the l_2-penalization. ENET LR refers to $0 < \alpha < 1$, although commonly $\alpha = 0.5$ is used. These solutions are a compromise from the sparse nature of LASSO solutions, and have improved predictive capabilities under high-dimensionality with correlated predictors (Zou and Hastie 2005).

The results for the three penalized methods using SLE FULL are presented in Table 1. Each method displays improved classification performance over LR. RIDGE, ENET, and LASSO produce mean out-of-set accuracies of 87.6%, 90.5%, and 89.2% for the unsmoothed curves, respectively. These values agree with the resulting accuracies

found from a recent investigation conducted by (Garbett and Brock 2016). In their study, a truncated temperature range of 60 – 82 ^{o}C was used to produce penalized models with out-of-set accuracies of 85 – 88%. Presented in this work are results using the full temperature range of 45 – 90 ^{o}C, which may explain the small increases to accuracy.

(Garbett and Brock 2016) did not include in their analyses derivative approximations of the SLE plasma thermograms. Using first derivative approximations of SLE FULL, mean test set accuracies are increased to 91.6% for RIDGE, 90.9% for ENET, and 90.0% for LASSO. The RIDGE results have mean sensitivity and specificity of 91.5% and 91.7%, indicating that both SLE and non-SLE are captured at similar rates. Second derivative approximations of SLE FULL show moderate classification performance, albeit reduced from first derivative results. These results provide evidence that inclusion of derivative information can aid disease classification using plasma thermograms.

Smoothing is found to have a slightly deleterious effect on the classification performance of predictive models for the original curves. Mean classification accuracy drops to 82.9% for RIDGE regression when using the GCV FULL predictors (Table 2). ENET and LASSO are similarly affected by the smoothing having mean test set accuracies of 85.5% and 85.3%. Models based on GCV derivative approximations display decreased performance, suggesting that smoothing the plasma thermograms has a deleterious effect on regression techniques. RIDGE, ENET and LASSO mean test accuracies drop to 87.7%, 87.5%, and 87.5%, respectively. Classification performance

using second derivative approximations is improved by smoothing, increasing mean accuracies to 86.1%, 86.6%, and 86.5%.

To evaluate how dimensionality of predictors influences SLE classification performance, the TRUNC discretization grid was used. The results of using SLE TRUNC and GCV TRUNC in the supervised learning algorithms are presented in Tables 3 and 4, respectively. Fewer predictors relieve the affects of collinearity on determining coefficients, but may cause the loss of important features. Classification performance of LR is improved, which for the TRUNC set results in 82.2% and 82.5% mean test set accuracies from original curves for the SLE and GCV sets. LR provides mean accuracies of 83.5/82.7% for first derivative and 78.1/82.0% for second derivative approximations using SLE TRUNC and GCV TRUNC. These results demonstrate that reducing the dimensionality of the problem and covariance of predictors improves LR, as expected.

The penalized methods, however, all have deleterious effects from the reduction of predictors using the TRUNC grid. RIDGE classification is the most influenced by reduced sampling from the functional approximations. RIDGE performance drops nearly 8% in mean accuracy from the FULL grid, resulting in 79.3% and 78.6% mean accuracies for the original curves from SLE TRUNC and GCV TRUNC, respectively. Derivative approximations have higher mean out-of-set accuracies than original curves under the TRUNC grid, but with losses from the FULL grid. First derivative approximations evaluated using RIDGE give mean test set accuracies of 85.4% and 84.3% using SLE TRUNC and GCV TRUNC. Second derivatives achieve 82.3% and 83.5%, confirming that smoothing does improve second derivative models. The reduced

Original Curves			
Method	Accuracy	Sensitivity	Specificity
LR	0.822 (0.049)	0.835 (0.068)	0.808 (0.072)
RIDGE	0.793 (0.052)	0.780 (0.075)	0.805 (0.075)
ENET	0.818 (0.050)	0.818 (0.070)	0.817 (0.072)
adap-ENET	0.809 (0.050)	0.811 (0.069)	0.808 (0.073)
LASSO	0.808 (0.051)	0.810 (0.072)	0.806 (0.074)
adap-LASSO	0.805 (0.053)	0.808 (0.072)	0.802 (0.075)
LDA	0.837 (0.045)	0.837 (0.063)	0.836 (0.067)
QDA	**0.873** (0.041)	**0.853** (0.061)	**0.894** (0.058)
KNN	0.763 (0.052)	0.725 (0.079)	0.803 (0.074)
First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.835 (0.047)	0.846 (0.063)	0.823 (0.067)
RIDGE	0.854 (0.040)	**0.864** (0.061)	0.843 (0.058)
ENET	0.854 (0.043)	0.859 (0.062)	0.849 (0.062)
adap-ENET	0.852 (0.046)	0.853 (0.065)	0.851 (0.065)
LASSO	0.853 (0.044)	0.851 (0.064)	0.855 (0.063)
adap-LASSO	0.855 (0.044)	0.854 (0.063)	0.855 (0.062)
LDA	0.846 (0.044)	0.861 (0.064)	0.831 (0.063)
QDA	**0.897** (0.035)	0.862 (0.058)	**0.934** (0.049)
KNN	0.893 (0.038)	0.916 (0.050)	0.870 (0.061)
Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.781 (0.051)	0.788 (0.072)	0.774 (0.079)
RIDGE	0.823 (0.048)	0.838 (0.065)	0.807 (0.075)
ENET	0.821 (0.050)	0.828 (0.069)	0.814 (0.073)
adap-ENET	0.810 (0.059)	0.818 (0.077)	0.801 (0.081)
LASSO	0.819 (0.050)	0.826 (0.071)	0.811 (0.072)
adap-LASSO	0.810 (0.051)	0.819 (0.072)	0.801 (0.073)
LDA	0.813 (0.046)	0.826 (0.066)	0.800 (0.076)
QDA	0.864 (0.044)	0.834 (0.061)	**0.895** (0.062)
KNN	**0.877** (0.041)	**0.922** (0.046)	0.830 (0.072)

Table 3. SLE TRUNC classification performance of the nine classifier as given by the accuracy, specificity, and sensitivity. Mean and standard deviation for each metric is given for the original curves and their first and second derivative approximations.

Original Curves			
Method	Accuracy	Sensitivity	Specificity
LR	0.825 (0.049)	0.829 (0.063)	0.820 (0.071)
RIDGE	0.786 (0.055)	0.780 (0.078)	0.792 (0.078)
ENET	0.835 (0.050)	0.840 (0.065)	0.830 (0.070)
adap-ENET	0.825 (0.051)	0.830 (0.067)	0.819 (0.074)
LASSO	0.825 (0.050)	0.830 (0.064)	0.818 (0.068)
adap-LASSO	0.823 (0.052)	0.830 (0.065)	0.816 (0.075)
LDA	0.843 (0.047)	0.838 (0.066)	0.847 (0.064)
QDA	**0.886** (0.037)	**0.866** (0.059)	**0.906** (0.054)
KNN	0.762 (0.052)	0.725 (0.079)	0.801 (0.075)
First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.827 (0.049)	0.840 (0.065)	0.813 (0.073)
RIDGE	0.843 (0.042)	0.839 (0.061)	0.848 (0.065)
ENET	0.850 (0.040)	0.853 (0.058)	0.848 (0.064)
adap-ENET	0.854 (0.040)	**0.864** (0.056)	0.843 (0.064)
LASSO	0.853 (0.040)	0.857 (0.055)	0.850 (0.065)
adap-LASSO	0.852 (0.041)	0.863 (0.056)	0.841 (0.066)
LDA	0.846 (0.044)	0.862 (0.061)	0.828 (0.068)
QDA	**0.887** (0.038)	0.855 (0.062)	**0.920** (0.052)
KNN	0.781 (0.052)	0.768 (0.074)	0.795 (0.074)
Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.820 (0.048)	0.827 (0.065)	0.813 (0.068)
RIDGE	0.835 (0.047)	0.834 (0.065)	0.836 (0.065)
ENET	0.831 (0.047)	0.826 (0.065)	0.837 (0.066)
adap-ENET	0.830 (0.047)	0.827 (0.068)	0.832 (0.067)
LASSO	0.831 (0.046)	0.829 (0.064)	0.833 (0.066)
adap-LASSO	0.832 (0.047)	0.829 (0.063)	0.834 (0.068)
LDA	0.841 (0.047)	0.839 (0.067)	0.842 (0.062)
QDA	**0.890** (0.037)	0.867 (0.059)	**0.914** (0.053)
KNN	0.805 (0.048)	**0.875** (0.058)	0.733 (0.089)

Table 4. GCV TRUNC classification performance of the nine classifier as given by the accuracy, specificity, and sensitivity. Mean and standard deviation for each metric is given for the original curves and their first and second derivative approximations.

dimensionality greatly impacts the predictive performance of RIDGE regression, which is reduced further by smoothing the functional representations. This result is not surprising as RIDGE regression performs best under high-dimensional settings where predictors may have strong correlation.

ENET and LASSO also show classification performance drops when using the TRUNC discretization. ENET mean test accuracies of 81.8% and 83.5% results from the original curves for SLE TRUNC and GCV TRUNC, while LASSO gives 80.8% and 82.5%. This results in an approximately 8% drop in classification accuracy using SLE FULL versus SLE TRUNC. Similar reductions in accuracy are found for first and second derivative approximations. The strength of ENET and LASSO regression is to perform simultaneous estimation of parameter selection and coefficient estimates. Reduction of the predictor set may have removed important predictors during the model building process, causing the loss of classification performance as observed.

Many of the penalized LR techniques also have adaptive variants. The adaptive methodologies refer to updated estimates of the shrinkage parameter λ after an initial regression solution has been achieved. For adaptive RIDGE, the resulting regression estimates have been shown to be identical to performing LASSO regression (Grandvalet 1998; Grandvalet and Canu 1999). The adaptive versions of LASSO (adap-LASSO) and ENET (adap-ENET) have the so called oracle property for conventional regression, implying that under asymptotic conditions the resulting regression models converge to the true underlying function which generates the observed responses (Zou 2006; Zou and Zhang 2009).

Adap-LASSO requires modifying the l_1-penalty term using the estimated coefficients from an initial LASSO estimation. Adap-ENET requires differential updates of the l_1- and l_2-penalty terms after initial estimates of ENET coefficients. The l_1-penalty terms are weighted by the ENET coefficients, while the l_2-penalties are retained from the initial ENET estimates (Zou and Zhang 2009). Although both procedures can be computationally expensive, adap-ENET requires significantly more computational sophistication. Adaptive updates should produce improved model performance due to the oracle property. However, under conditions where predictors have high multicollinearity, predictive performance of the adaptive models can be reduced from the original estimated models (Chan and Chen 2011).

Tables 1 and 2 demonstrate clearly that adaptive penalized LR estimators return unfavorable models for the FULL discretization grid. The SLE FULL classification performance of adap-LASSO and adap-ENET are reduced to 87.4% and 86.4% from the results found prior to the adaptive updates. First and second derivative based adaptive models also have decreased classification performance. Smoothing of the functional approximations also leads to losses in classification accuracy after adaptive updates, although the reduction in accuracy is less severe. Thus, adaptive strategies for penalized LR all fail to improve classification performance under the FULL discretization.

The reduced dimension predictor sets (Tables 3 – 4) show less effect of the adaptive models on LR performance, occasionally improving mean test set accuracy. These results are believed to relate to predictor collinearity, which causes significant inflation of variance that inhibits adaptive updates from converging properly. This

accounts for the loss in mean test set accuracy for adaptive solutions under the FULL discretization. Collinearity is relieved under the TRUNC discretization and adaptive performance improves. Although adaptive models still generally show decreased mean test set accuracies, the severity of the decrease is diminished. Several adaptive models even improve performance for derivative-based classification models.

The final regression method investigated was least squares approximations (LSA) to adaptive-LASSO. Postulated by (Wang and Leng 2007) is that asymptotically the l_1-penalization problem can be rewritten for LASSO estimation as

$$(\beta - \tilde{\beta})' \hat{\Sigma}^{-1} (\beta - \tilde{\beta}) + \lambda \sum_{j} |\beta_j|.$$

The concept being that the loss-function can be asymptotically approximated using a consistent covariance matrix $\hat{\Sigma}$ and estimated coefficients $\tilde{\beta}$. Both estimated coefficients and the covariance matrix are standard output of many R functions. LSA is capable of producing adap-LASSO estimates requiring only a single LR fit. This greatly simplifies the computational needs of the adaptive methodologies. A primary LR analysis is performed, and the resulting covariance matrix and estimated coefficients are used to produce estimates of adap-LASSO coefficients. The algorithm produced by (Wang and Leng 2007) provides coefficients based on iterative updates of the penalization parameter λ using either Akaike or Bayes information criteria.

LSA clarifies the difficulties of multicollinearity between the predictors of the SLE plasma thermogram data set: it produces models which are all null (results not shown) and have correspondingly low performance. LSA is incapable of converging to

the effective estimates of the adap-LASSO solution because the resulting covariance matrix $\widehat{\Sigma}$ is nearly singular. This is the problem inherent to the adaptive methods studied, and explains why adaptive solutions have decreased classification performance for the SLE plasma thermograms.

The results presented here are the first in-depth investigation of predictive regression models built from derivative approximations of the SLE plasma thermograms. The investigation conducted demonstrates that derivatives approximations of the SLE plasma thermograms can produce classifiers whose out-of-set accuracies outperform using only original curve information. RIDGE results in 91.6% mean test set accuracy when unsmoothed first derivative approximations were evaluated. This represents a classification performance higher than all currently published studies, and does so with both high sensitivity (91.5%) and specificity (91.7%). This motivates the use of derivatives in developing classification models for SLE plasma thermograms, while also suggesting that models which combine information from multiple derivatives may have strong increases in performance. The uses of ensemble methodologies are to be explored in Chapter 5.

4.3 Discriminant Analysis

Discriminant analysis takes an alternative approach to classification of responses than the regression techniques evaluated above. LR directly evaluated the probability that a particular observation belongs to the kth class given the predictor information (i.e. $P(Y = k|X)$. Bayes' theorem gives us that

$$P(Y = k|X) = \frac{P(X|Y = k)P(Y = k)}{P(X)}.$$

Discriminant analysis uses Bayes' theorem to instead evaluate how likely an observation came from a class based on the density functions of each class (James et al. 2013). For the binary case, conditional probability density functions for $P(X|Y = 0)$ and $P(X|Y = 1)$ are assumed to be multivariate normal distributions with mean and covariance $(\boldsymbol{\mu_0}, \boldsymbol{\Sigma}_0)$, $(\boldsymbol{\mu_1}, \boldsymbol{\Sigma}_1)$, respectively. The parameters of the distributions are estimated from the observations within each class, with the prior probability $P(Y = k)$ being the prevalence of the kth class. LDA makes the additional assumption of homoscedasticity in the class covariance giving $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$. This allows for the log-probability of an observation being of the kth class can be written as

$$\log\big(P(Y = k|X)\big) = X\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k} - \frac{1}{2}\boldsymbol{\mu_k}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k} + \log\big(P(Y = k)\big).$$

Decisions are then drawn from log-probabilities that are linear combinations of the known observations. An out-of-set decision is made by choosing the class with the largest log-probability.

Supervised learning based on LDA was conducted using the **MASS** package of R. The classification of SLE vs. non-SLE alternatives was evaluating using SLE FULL, GCV FULL, SLE TRUNC, and GCV TRUNC show in Tables 1 – 4. LDA has improved classification performance when based on GCV functional approximations. GCV FULL produces mean test set accuracies of 85.3%, 85.4%, and 85.6% for the original curves

along with first and second derivative approximations. This is in contrast to the unsmoothed results of 74.0%, 74.1%, and 74.3%.

LDA is equally affected by the choice of discretization grid. Using the TRUNC grid, unsmoothed predictors improve in classification performance over the FULL grid. Mean test set accuracies are increased to 83.7%, 84.6%, and 81.3% for SLE TRUNC original curves, first, and second derivatives. This is likely a consequence of improved estimation of covariance due to alleviation of predictor collinearity. GCV TRUNC produces reduced mean accuracies of 84.3%, 84.6%, and 84.1% from the GCV FULL results. Reduction of the predictor set may cause significant predictors to be removed, which could explain the minor losses in classification accuracy. LDA shows similar trends to the regression methods in that models based on derivative approximations improve upon using smoothed functions. For the FULL discretization grid, first derivative approximations are improved nearly 10% upon the introduction of smoothing using a reduced basis expansion.

LDA offers a computationally fast classification method that does not require complex minimization routines, providing efficient results. The penalized methods outperform LDA for nearly all investigated sets, but do so at the cost of computational cost and complexity. To improve the discriminant methods, the assumption of homoscedasticity can be dropped, allowing for the covariance of each conditional probability to be different. With unequal covariance, the log-probability for an observation being of the kth class becomes the quadratic function

$$\log\big(P(Y = k|X)\big) = -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu_k})'\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{X} - \boldsymbol{\mu_k}) - \frac{1}{2}\log|\boldsymbol{\Sigma_k}| + \log\big(P(Y = k)\big).$$

These log probabilities produce QDA, a more flexible alternative to LDA that requires estimation of covariance matrices for each conditional probability function.

QDA estimation can be problematic due to collinearity of predictors, like that seen with penalized LR. QDA could only be performed on the TRUNC discretization grid, with singularity of the covariance estimates inhibiting the use of the FULL predictors. Resulting classification performances for QDA are shown in Tables 3 – 4. QDA results in mean test set accuracies of 87.3% and 88.6% for SLE TRUNC and GCV TRUNC, second only to ENET and LASSO models performed on the FULL grid. QDA is more sensitive to the smoothing of the predictors than LDA, likely due to the quadratic nature of the decision boundary. QDA produces excellent classification performance for derivative approximations as well. Mean test set accuracies of 89.7% and 86.4% result from analysis of the first and second derivative approximations of SLE TRUNC. GCV TRUNC gives mean accuracies of 88.7% and 89.0%, further illustrating that smoothing improves the predictive performance of second derivative models.

The discriminant methods have been shown to be effective methods for SLE plasma thermogram classification. LDA has been established as an excellent classification method for SLE plasma thermograms in recent studies (Garbett and Brock 2016; Garbett et al. 2017). In the more recent study, a modified LDA approach was used in combination with predictors from anti-body screening to produce a mean classification accuracy of 89%. Their work exemplifies the potential for combining models that evaluate thermogram data along with patient, serological, and immunological predictors.

This dissertation included an evaluation of QDA, which allows for more flexible parameter selection. The quadratic nature of the decision boundary could have benefits for capturing SLE vs. non-SLE cases. QDA did require reduction of the predictor set, with problems inverting the covariance matrix occurring on the FULL predictor mesh. QDA is well-posed on the TRUNC predictor mesh, returning mean accuracies second only to the penalized LR methods using unsmoothed first derivative approximations. Smoothing did have a minor influence on second derivative QDA models returning small classification gains. QDA produced high classification accuracies for the original curves and derivative approximations; unique to QDA was that high accuracy models were obtained through high specificity. QDA based on unsmoothed first derivative models resulted in a mean specificity of 93.4%, the highest of all implementations considered. With the success of both original curves and derivative based models, discriminant methods could be effective under ensemble methodologies that are of future research interest to the author.

4.4 K-Nearest Neighbors

The final contemporary method considered was the nonparametric method KNN. In KNN a measure of distance is defined between observations; the metric is then used to predict the class of an unknown observation based on the known classes of its K-nearest neighbors. Let N_0 be the set of training observations that are closest to the test observation. The conditional probability that the test observation is of the kth class is then defined as

$$P(Y = k|X) = \frac{1}{K}\sum_{j \in N_0} I(y_j = k).$$

where K is the number of neighbors considered. KNN assigns conditional probabilities by simply counting the classes of its K neighbors.

The method is sensitive to the tuning constant K. Decision boundaries can be overly flexible when K is small and become inflexible for large K. The parameter is commonly validated during supervised learning. KNN also relies on the metric used to determine the distance between observations. The most commonly used metric in $\mathbb{R}^p$ is the Euclidean norm, which defines the distance $d(\boldsymbol{X}_1, \boldsymbol{X}_2)$ between two predictor vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ by

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) = \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| = \sqrt{\sum_j (x_{1,j} - x_{2,j})^2}.$$

KNN can be adapted to a variety of norms, and on occasion selection of the norm can influence classification performance.

Base R package **class** can be used to perform KNN using the Euclidean norm discussed above. The results of KNN on the four predictor sets are summarized in Tables 1 – 4. Cross-validation of the neighbor size, K, was performed during the supervised learning algorithm for each implementation. KNN returns subpar classification performance for the evaluation of original curves from each of the four discretization grids, with a mean test set accuracy of 76.2%. Derivative approximations show promising results for KNN implementation though. SLE FULL results in a mean test set

accuracy of 90.8%, which is competitive among all methods tested, and returns the highest mean test set sensitivity of 94.5%. High mean test set sensitivities are found for several of the KNN implementations, commonly having sensitivities higher than 90%.

Smoothing and reduced sampling both hinder classification performance of KNN. A loss in predictive performance is observed for each of the GCV FULL, SLE TRUNC, and GCV TRUNC implementations compared to SLE FULL. KNN models based on first derivatives return only 77.4% mean test set accuracy, a drop of 13% from the unsmoothed curves. Second derivative approximations between SLE FULL and GCV FULL are nearly equivalent. This is commonly observed from many of the tested methods, with smoothing aiding the evaluation of second derivative approximations. Use of the SLE TRUNC predictor set drops first derivative KNN mean test set accuracy to 89.3%. Although not as drastic a change as smoothing, less sampling from the latent functions does correlate with a loss in KNN predictive performance. KNN results indicate the potential for producing high accuracy and high sensitivity classifiers based on derivative approximations, with performance competitive with penalized LR.

4.5 Combined Predictor Matrices

The combination of predictors from multiple derivative orders was studied next. Specifically, the discretized predictors from original curves were combined with first derivative approximations or with both first derivative and second derivative approximations. The combined predictor matrices were then evaluated using the same

nine classifiers discussed above. This provides an evaluation of how derivative-based predictors can influence classification results in combination with the primary curve predictors.

The two discretization grids lead to significantly different sets of predictors. Under the FULL discretization, 451 predictors are sampled from the temperature range 45 – 90 ^{o}C. This produces matrices with 902 and 1353 predictors to be analyzed against a total of 589 patient samples. Therefore, under the FULL discretization, both the combination of original curve predictors with first derivative approximations, and the original curves combined with first and second derivative approximations produce high-dimensional problems (i.e. more predictors than samples). The TRUNC grid was also evaluated, returning 91 predictors per curve. Even with the combination of all three curves, only 293 predictors are used and the classification problems are still over-determined. These results continue to show sub-par classification performance in comparison with the FULL grid, in agreement with findings in Section 4.2. The resulting classification performances using the combined predictors from SLE TRUNC and GCV TRUNC are presented for reference in Appendix A, Tables A1 and A2.

The results of supervised classification using the combined predictor matrices for the SLE FULL discretization are provided in Table 5. LR failed to converge due to the high-dimensional setting and is referred to as under-determined (UD). Additionally, QDA requires lower dimensional settings and reduced predictor collinearity and failed to converge for all sets investigated, save the SLE TRUNC original and first derivative combination (Appendix A, Table A1). QDA performance under this setting was

Original + First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	UD	UD	UD
RIDGE	0.849 (0.049)	0.834 (0.069)	0.865 (0.063)
ENET	**0.911** (0.038)	**0.915** (0.054)	**0.908** (0.054)
adap-ENET	0.886 (0.045)	0.894 (0.059)	0.878 (0.067)
LASSO	0.899 (0.039)	0.905 (0.055)	0.894 (0.057)
adap-LASSO	0.878 (0.046)	0.889 (0.06)	0.867 (0.068)
LDA	0.740 (0.056)	0.756 (0.083)	0.724 (0.081)
QDA	DNC	DNC	DNC
KNN	0.830 (0.047)	0.829 (0.067)	0.832 (0.066)
Original + First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	UD	UD	UD
RIDGE	0.887 (0.039)	0.881 (0.058)	0.894 (0.056)
ENET	**0.907** (0.036)	0.909 (0.051)	**0.905** (0.051)
adap-ENET	0.867 (0.045)	0.869 (0.062)	0.865 (0.07)
LASSO	0.897 (0.037)	0.9 (0.054)	0.893 (0.052)
adap-LASSO	0.853 (0.046)	0.859 (0.061)	0.848 (0.072)
LDA	0.740 (0.065)	0.738 (0.113)	0.742 (0.050)
QDA	DNC	DNC	DNC
KNN	0.883 (0.041)	**0.937** (0.042)	0.829 (0.075)

Table 5. SLE FULL combined predictor classification performance of the nine classifiers as given by the accuracy, specificity, and sensitivity. Predictors were produced by combining discretized predictors from original curves with first derivative (Original + First Derivative) or with both first and second derivative approximations (Original + First Derivative + Second Derivative). The test set mean and standard deviation for each metric is recorded. UD represents under-determined solutions, while DNC represents solutions that did not converge.

equivalent to the results found analyzing only SLE TRUNC first derivative approximations, thus the combination of predictors seemed to have minor influence on QDA.

Table 5 demonstrates that combined predictors across multiple derivative orders can influence SLE classification performance. SLE FULL original curves combined with first derivative predictors produces improved performance for ENET LR. The mean out-of-set accuracy increases to 91.1% for ENET for this combination. This result is slightly reduced to 90.7% when the second derivative predictors are also included. These results display that ENET is capable of performing well under high-dimensional settings and under conditions of high multicollinearity. The 91.1% mean test set accuracy is second only to the performance of RIDGE LR using only the first derivative approximations (Table 1).

LASSO performance is nearly equivalent when using first derivative predictors (Table 1) or a combination of original and first derivative predictors (Table 5). This suggests that the LASSO solutions continue to capture equivalent models even under high-dimensional settings. This is an established property of LASSO methods, which continues to hold even when all three discretized predictors matrices are combined. LASSO produces mean test set accuracies of 90.0%, 89.9%, and 89.7% using original curve predictors, original with first derivative predictors, and original with first and second derivative predictors, respectively. Thus, LASSO models have a high degree of reproducibility even when combining predictors from multiple derivative orders.

RIDGE is most influenced by the combination of predictors. When using original curve predictors in combination with first derivative predictors, RIDGE classification performance drops to a mean test set accuracy of only 84.9%. This result is clearly diminished from the models produced using only first derivative predictors, which returned the highest mean test set accuracy of all methods, 91.6%. RIDGE performance rebounds slightly when second derivative are also included in the combined matrix, provided a mean accuracy of 88.7%, still reduced from first derivative predictors alone. RIDGE LR is repressed by increased predictor matrices, indicating that regularization of the linear model coefficients is not enough to produce satisfactory models under high-dimensional settings. The predictor selection properties of ENET and LASSO are influential at producing useful classifiers from the high dimensional SLE predictor sets. However, it is clear that RIDGE can produce high performance models when used on only the first derivative predictors.

The outcomes of adaptive penalized strategies continue to show a loss in predictive performance, similar to the findings from Section 4.2. Both adap-ENET and adap-LASSO show dips in mean test set accuracies, with losses of nearly 3 – 4% accuracy after the adaptive updates. This continues to be related to high multicollinearity of the predictors, which is still present under the conditions of the combined predictor matrices.

Table 5 also depicts classification performance for LDA and KNN. LDA has considerable issues with multicollinearity under the high-dimensional settings. LDA solutions do converge, but provide only 74.0% mean test set accuracy for each combined

predictor set. LDA performance is significantly improved when smoothing using B-spline basis reduction is introduced. These results are shown in Table 6, where the supervised classification performance of all nine methodologies is shown for GCV FULL predictor combinations. Although all other classifiers have reduced mean test set accuracy upon smoothing of the curves, LDA performance is increased to 85.3%. This nearly 10% increase in classification performance is consistent whether only single predictor matrices are used, or if predictors from multiple derivatives are combined. This suggests that LDA is more sensitive to the oscillations within the raw data observations, and that minimal smoothing such as that introduced by B-spline basis reductions can drastically influence LDA performance.

The final model building strategy, KNN, shows unique results when using combined predictor matrices. KNN had nearly equivalent classification performance to penalized LR when using only first derivative approximations (Table 1). KNN performance is repressed by the combination of first derivative predictors with original curve predictors, dropping mean test set accuracy to 83.0% from 90.8%. However, much of the predictive performance can be regained if second derivative predictors are also introduced. Original curves combined with first and second derivative predictors provide a mean test set accuracy of 88.3%. This increase in performance, from 83.0% to 88.3%, upon introduction of second derivative predictors, occurs almost solely through improved sensitivity in the model performance. KNN continues to produce models with the highest sensitivity to SLE, with the original curves combined with first and second derivative predictors giving a mean test set sensitivity of 93.7%. This result suggests that KNN, and

Original + First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	UD	UD	UD
RIDGE	0.806 (0.052)	0.795 (0.074)	0.817 (0.076)
ENET	**0.877** (0.041)	0.878 (0.057)	**0.875** (0.063)
adap-ENET	0.848 (0.052)	0.853 (0.066)	0.842 (0.076)
LASSO	0.875 (0.041)	**0.879** (0.055)	0.87 (0.063)
adap-LASSO	0.863 (0.046)	0.866 (0.059)	0.86 (0.066)
LDA	0.853 (0.042)	0.867 (0.060)	0.838 (0.065)
QDA	DNC	DNC	DNC
KNN	0.768 (0.054)	0.731 (0.082)	0.805 (0.075)
Original + First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	UD	UD	UD
RIDGE	0.847 (0.046)	0.850 (0.066)	0.844 (0.072)
ENET	0.877 (0.040)	**0.877** (0.057)	0.876 (0.060)
adap-ENET	0.844 (0.047)	0.846 (0.063)	0.842 (0.073)
LASSO	0.873 (0.041)	0.875 (0.057)	0.871 (0.063)
adap-LASSO	0.850 (0.044)	0.847 (0.065)	0.852 (0.071)
LDA	0.853 (0.042)	0.867 (0.060)	0.838 (0.065)
QDA	DNC	DNC	DNC
KNN	**0.888** (0.035)	0.869 (0.057)	**0.907** (0.051)

Table 6. GCV FULL combined predictor classification performance of the nine classifiers as given by the accuracy, specificity, and sensitivity. Predictors were produced by combining discretized samples from original curves with first derivative (Original + First Derivative) or with both first and second derivative approximations (Original + First Derivative + Second Derivative). The test set mean and standard deviation for each metric is recorded. UD represents under-determined solutions, while DNC represents solutions that did not converge.

potentially other nonparametric classification algorithms, can produce models whose sensitivities are far higher than the other model building strategies.

Combined predictor matrices show the capabilities of combining information across derivative orders. Although ineffective for most supervised learning algorithms investigated in this work, the results show potential for improving classification performance by using a mixture of information from multiple derivative orders. This naive approach addressed only if simple mixtures of predictors could influence classification performance. Ensemble strategies will study how classifiers produced using predictors from different curves (i.e. original, first derivative, second derivative) can be combined to improve performance. Ensemble strategies will be introduced in Section 5.6, and can be used for either discretized predictors or functional data based classifiers.

4.6 Conclusions

Supervised classification of SLE plasma thermograms is improved by including derivative approximations as predictors. FDA B-spline basis functional representations allow for efficient discretizations of original curves and their derivative approximations. Contemporary model building strategies applied to the original curves return solutions matching recent publications (Garbett and Brock 2016). Classification performance of penalized LR returns accuracies nominally 86.4 – 90.5%. These mean test set accuracies are slightly higher than the 85 – 88% observed in previous research efforts, but are likely caused by differences in the range of temperatures included in the analysis.

Large improvements in classification performance are observed for the first derivative approximations. RIDGE LR returns KCV mean test set accuracies of 91.6%, higher than any previously published result. The resulting classifiers achieve high accuracy with both high sensitivity (91.5%) to SLE and high specificity (91.7%) indicating that false negatives are reduced. ENET and LASSO also produce 90% or higher mean test set accuracies for predictors based on first derivatives.

NC using KNN has sub-par performance when using original curve discretizations. KNN performance peaks when using first derivative predictors, returning a KCV mean test set error of 90.8%. KNN produces the highest mean test set sensitivity of 94.5%, suggesting that the method is capable of capturing SLE cases with high success, although at the expense of increased false negatives. The sensitivity to SLE cases could be a useful property of nonparametric methods, if only potential diagnosis of the disease is desired. If SLE plasma thermograms are to be used as a fast and efficient screening measure, then high sensitivity to SLE cases may be desirable.

FDA was also used to derive smoothed functional representations. Smoothing of the B-spline functions through basis reduction produced losses in classification performance of the penalized methods. Discriminant-based classification is improved, but the models are significantly reduced from penalized LR of unsmoothed approximations. QDA has surprisingly strong performance, but requires that the predictor set be truncated to one-fourth the size of the raw data output.

These results suggest that unsmoothed functional representations should be the primary source of predictors, with the most effective classifiers being based on derivative

predictors. Reduction of the predictor dimensionality is nominally deleterious. Methods to improve how the predictors are sampled from the functional representations could offer further improvements (Berrendero et al. 2016; Delaigle et al. 2012), and are a source of interest for future work.

Combining predictors from multiple derivative approximations produces only minor changes to classification performance. Many of the methods return small losses in accuracy due to the high dimensionality of the problem and collinearity of predictors. When combining original curves with first and second derivatives, 1353 predictors are produced with only 589 patients available for classification; this represents a high-dimensionality classification problem causing several of the methods to be ineffective. Penalized LR works well in these situations, with ENET returning a mean test set accuracy of 91.1% when the original curve and first derivative predictors are combined. However, such models are computationally intensive, with results of RIDGE regression on just first derivative approximations still returning the highest mean accuracy.

Combining predictors does suggest that information shared across multiple curves could boost classification performance. The ENET solution of 91.1% does improve over the solution for either of the curves alone. Selection of predictors using sure information screening could improve results from combined predictor matrices (Saldana and Feng 2016), and is of interest for future studies. More sophisticated methods for combining information could also produce better gains in classification. Chapter 5 will introduce ensemble strategies that will allow the predictions from multiple models to be combined, which will be shown to have improvements to predictions.

In this chapter, FDA was only used to return discretized predictors, and was an effective method for dealing with derivatives. Potential improvements to the identification of SLE plasma thermograms using functional classification will be investigated in Chapter 5. This chapter currently represents the most detailed supervised classification study of SLE plasma thermograms using contemporary methods. Resulting models from derivative approximations have been shown to be capable of producing higher accuracies than previous investigations: the presented results should motivate investigators to include derivative information in future plasma thermogram studies.

Chapter 5

FUNCTIONAL SUPERVISED CLASSIFICATION AND ENSEMBLE STRATEGIES

5.1 Introduction

Studied thus far were contemporary strategies based on the classical concept of discretized predictors, generated by sampling the latent functions, $X(t)$, which represent the SLE plasma thermogram signatures. The predictors were then introduced to classic model building strategies: LR, discriminant analysis, and KNN. Instead of discretizing the latent functions, FDA classification can be performed utilizing the functional representations (Ferraty and Vieu 2006; Ramsay and Silverman 2007). FDA classification has not been previously evaluated for its performance in predicting SLE vs. non-SLE cases. Introduced here will be three functional analogues of LR: functional generalized linear models (FGLM), functional generalized spectral additive models (FGSAM), and functional generalized kernel additive models (FGKAM).

FuNC will be introduced through functional KNN classifiers. FuNC uses integration to estimate distances between curves and serve as the core classifiers used in the algorithm developed in Chapter 6. Ensemble learning strategies that combine probabilities from multiple models will then be explored. Several types of ensembles will be investigated including naïve voting using predicted classes and weighted voting of prediction probabilities. Ensemble learning using the resulting models developed in

Chapter 3 will be evaluated. This will allow predictions to be made based on multiple model outputs rather than attempting to build models from enlarged predictor sets. Ensemble learning increases overall test set accuracies for all methodologies studied.

5.2 Functional Logistic Regression

LR is a GLM under the binomial distribution with the logit link function. This produces log-probability estimates which are linear combinations of the predictors, introduced in Section 4.2. Under the functional perspective, the linear combination of covariates is replaced by the inner product in the functional space (James 2002). Where GLMs evaluate the linear combination of discrete covariates, FGLMs consider the inner product over the functional space (Müller and Stadtmüller 2005). Consider the FDO comprised of n random functional variables, $X_i(t)$, defined on the support T each associated with a binary response variable y_i. The probability that the ith observation belongs to class 1 is then given as

$$p(X_i) = P(Y = 1|X_i(t): t \, \epsilon \, T) = \frac{\exp\{\beta_0 + \int_T X_i(t)\beta(t)\, dt\}}{1 + \exp\{\beta_0 + \int_T X_i(t)\beta(t)\, dt\}}$$

with β_0 a real parameter and $\beta(t)$ a parameter function (Escabias et al. 2004). The logit transformation produces the corresponding functional LR (FLR) model

$$\log\left(\frac{p(X_i)}{1 - p(X_i)}\right) = \beta_0 + \int_T X_i(t)\, \beta(t) dt, \qquad i = 1, \dots, n.$$

Solutions to the above FGLM are well studied in the literature, with FLR having been incorporated into several R packages (Febrero-Bande and de la Fuente 2012; Müller and Yao 2008). The R package **fda.usc** provides a wide set of functional classification tools and was used for SLE functional regression. R functions allow for a mixture of functional covariates and conventional covariates. Let $\pi_i = p(X_i)$ be the predicted class probability of the ith subject based on functional and non-functional covariates.

The most fundamental solution builds upon basis expansions of the functional representations (Ratcliffe et al. 2002). Let the linear expansions of $X(t) = \mathrm{c_i}\varphi(\mathrm{t})$ and $\beta(t) = \psi^{\mathrm{T}}(\mathrm{t})\mathrm{b}$. Then the FLR model can be rewritten

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = z_i^T\alpha + \int_T \mathrm{c_i}\varphi(\mathrm{t})\psi^{\mathrm{T}}(\mathrm{t})\mathrm{b}\,dt.$$

The conventional covariates of the ith subject are given as $z_i^T = [1\ z_1\ \ldots\ z_r]$ with parameters $\alpha = [\alpha_0\ \alpha_1\ \ldots\ \alpha_r]^T$.

Basis expansions must be chosen carefully but improve the estimation of FLR models. Let $\mathrm{W} = \int \varphi(\mathrm{s})\psi^{\mathrm{T}}(\mathrm{s})\,ds$ and rewriting into matrix notation produces

$$\log\left(\frac{\pi}{1-\pi}\right) = \mathrm{Z}\alpha + CWb = [Z\ CW]\begin{bmatrix}\alpha\\ b\end{bmatrix}.$$

This gives the FLR model in a form similar to standard LR models. ML estimates can then be found using Fisher scoring and the number of basis functions determined by cross-validation.

Models that use a combination of functional and non-functional covariates could allow for patient information as well as serological and immunological predictors to be

combined with the FDOs during FLR. Methodologies that allow a combination of predictors across diagnostic tests are of future research interest. Such models are responsible for recent publications providing the current top classification for SLE plasma thermograms (Garbett et al. 2017), although these models only produce 89% mean test set accuracies.

Alternative methods for estimating FLR using FPCA have been introduced by (Escabias et al. 2004). The FPCA FLR models do not require estimation of standard LR coefficients that produce highly correlated covariates. The method relies on two distinct approximation techniques for analyzing FPCA. A reduced set of FPCs are used as covariates producing a FGLM that evaluates the inner product of basis functions with reduced correlation; this stabilizes the computational estimates and improves classification accuracy.

FLR classification of the SLE plasma thermograms was investigated using FGLM functions developed in **fda.usc**. The functional representations SLE FDO and GCV FDO were used as functional covariates. FGLM models were estimated for original, first derivative, and second derivative curves with classification results given in Table 7. Using the basis expansion method for estimating FGLM produces virtually equivalent results to contemporary classification as expected from the discussion above. A mean test set accuracy of 71.6% results from FGLM models of original curves in the SLE FDO, in excellent agreement with the 72.0% from traditional analysis. Functional classification based on first derivative approximations are slightly increased, with FGLM

Original Curves			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.716 (0.059)	0.716 (0.08)	0.715 (0.084)
FGSAM	0.740 (0.055)	0.716 (0.076)	0.765 (0.082)
FGKAM	0.723 (0.057)	0.672 (0.086)	0.774 (0.080)
FKNN	**0.763** (0.052)	**0.722** (0.079)	**0.804** (0.074)
First Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.736 (0.061)	0.711 (0.083)	0.761 (0.084)
FGSAM	0.761 (0.053)	0.752 (0.074)	0.771 (0.074)
FGKAM	0.748 (0.056)	0.773 (0.076)	0.722 (0.080)
FKNN	**0.905** (0.040)	**0.928** (0.047)	**0.881** (0.063)
Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.715 (0.059)	0.697 (0.09)	0.734 (0.083)
FGSAM	0.728 (0.057)	0.685 (0.092)	0.772 (0.081)
FGKAM	0.836 (0.045)	**0.977** (0.026)	0.691 (0.090)
FKNN	**0.882** (0.041)	0.936 (0.044)	**0.826** (0.071)

Table 7. Functional classification results using SLE FDO. Classification performance is summarized by accuracy, specificity, and sensitivity. The mean and standard deviation for each metric are presented.

producing a mean test set accuracy of 73.6% compared to 71.8% from classic methods. Second derivative FGLM models return 71.5% mean test set accuracy.

Smoothing the functional approximations using B-spline basis reduction was investigated in Table 8. Functional classification using the GCV FDO is nearly unchanged from the SLE FDO, with classification from original curves reproducing the 71.6% mean test set accuracy. First derivative mean test accuracy is slightly improved to 74.7%, with second derivatives mimicking unsmoothed results, giving 71.5%. This suggests that smoothing has less influence on functional classification than contemporary LR. Smoothing caused a large gain in classification accuracy for the discretized sets, with mean test set accuracies increasing on average more than 5%. Overall, FLR based on FGLM has a classification performance similar to unsmoothed discretized evaluations, producing mean test set accuracies of 70 – 75%.

5.3 Functional Generalized Additive Models

A second approach to the solution of the FLR approaches through generalized additive models (GAM). Introduced by (Hastie and Tibshirani 1990), GAMs allow linear predictors to depend on smooth functions of the predictors. Instead of a linear combination of explanatory variables, such as in GLMs, the estimator is produced through the linear combination of smoothed functions. There are several methods for estimating the smoothing functions using local likelihood, spline expansion, PCA, and kernel smoothing. A functional GAM (FGAM) model for FLR can be expressed as

Original Curves			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.716 (0.059)	0.716 (0.08)	0.715 (0.084)
FGSAM	0.740 (0.055)	0.716 (0.076)	0.765 (0.082)
FGKAM	0.723 (0.057)	0.673 (0.086)	0.774 (0.080)
FKNN	**0.762** (0.052)	**0.722** (0.079)	**0.802** (0.075)
First Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.747 (0.057)	0.723 (0.08)	0.773 (0.079)
FGSAM	**0.775** (0.051)	0.759 (0.076)	**0.790** (0.074)
FGKAM	0.725 (0.057)	0.713 (0.083)	0.737 (0.081)
FKNN	0.772 (0.051)	**0.769** (0.073)	0.774 (0.075)
Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.715 (0.057)	0.699 (0.086)	0.731 (0.081)
FGSAM	0.735 (0.053)	0.696 (0.080)	0.775 (0.076)
FGKAM	0.767 (0.052)	**0.952** (0.038)	0.578 (0.098)
FKNN	**0.874** (0.042)	0.940 (0.042)	**0.807** (0.076)

Table 8. Functional classification results using GCV FDO. Classification performance is summarized by accuracy, specificity, and sensitivity. The mean and standard deviation for each metric are presented.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$$

where $f_j(X_j)$ are estimations of partial smoothing functions (Febrero-Bande and González-Manteiga 2013), and X_j the sample functional covariates.

The difficulty with the FGAM above is the estimation of the partial functions. One solution is use of spectral decompositions based on FPCA. FPCs are utilized in an additive rather than linear way as discussed in Section 5.2. The uncorrelated FPCs are used as predictors for the partial smoothing functions, which can be estimated using one-dimensional smoothing steps. Each $f_j(\cdot)$ is estimated by local linear regression using the kth FPC scores and does not require back-fitting (Müller and Yao 2008). This method is termed FGSAM with functions available through the **fda.usc** R package. Supervised learning algorithms were developed to evaluate how FGSAM classification improved while increasing the number of FPC included in the analysis. KCV was used on a grid of FPC component sizes, where it was found that inclusion of the first four FPC produced optimized models for each set of curves investigated.

The classification performance of FGSAM using SLE FDO and GCV FDO are depicted in Tables 7 and 8, respectively. Classification performance is improved from FLR, but with mean test set accuracies still well below the penalized regression methods of Chapter 4. Mean accuracies from KCV for the original, first derivative, and second derivative functions are 74.0%, 76.1%, and 72.8%. This shows that first derivative curves continue to provide improved classification performance, but that FGSAM only slightly improve overall performance. Using reduced basis expansions for the FPCA

does produce slight increases in classification performance of derivative curves: this agrees with all previous results that smoothing improves derivative models.

A second FGAM estimates the partial smoothing functions through functional nonparametric kernel estimators (Ferraty and Vieu 2006). Introduced by (Febrero-Bande and González-Manteiga 2013) are back-fitting algorithms that allow for nonparametric estimation of the partial functions. The algorithm produces updates to the smoothing functions using asymmetric kernel functions. The methodology has improved properties for the convergence of the smoothing solutions, uses only distances between covariates, can use any L_p-metric, and has reduced effects due to curse of dimensionality. Back-fitting and kernel estimation have computational costs, causing the method to be slower than FGSAM.

Functions for estimating FLR classifiers using FGKAM are implemented in **fda.usc**. The results of FGKAM for SLE FDO and GCV FDO are shown in Tables 7 and 8. The kernel-based functional models have similar classification performance to other functional methods for the original and first derivative curves. Of unique interest is the large improvement in classification for the second derivative models. Mean test set accuracy for SLE FDO second derivatives increases to 83.6% with a 97.7% sensitivity, indicating that SLE cases are almost always correctly classified. Such results mimic the KNN results presented in Chapter 4, which produced high prediction accuracies through high sensitivity models. The kernel-based functional models using second derivative curves show a similar property. The GCV FDO returns a mean test set sensitivity of 95.2% for second derivative models, but mean test set accuracy drops to 76.7%. The

FGKAM models show similar resulting classification performance as nonparametric estimators in that sensitivity is high, countered by low specificity.

5.4 Functional K-Nearest Neighbors

The final model building strategy is based on nonparametric FDA (Ferraty and Vieu 2006). FuNC redefines the concept of closeness by allowing for the Euclidean norm used in contemporary KNN to be replaced by the L_2-metric. The distance $d(X_i, X_j)$ is defined for any two functional covariates, $X_i(t)$ and $X_j(t)$, by

$$d(X_i, X_j) = \sqrt{\int \left(X_i(t) - X_j(t)\right)^2 dt}.$$

This provides a flexible framework for generating distance approximations, and under the functional setting is less affected by the curse of dimensionality. Sampling from the functions now relates to integration accuracy, and can be controlled easily based on the desired computational needs. FuNC allows any L_p-metric to be used to produce distance estimates, although primarily the L_2-metric is utilized.

The next two chapters focus on FuNC based learning algorithms. The proposed algorithms implement distance calculations performed using Simpson's rule to estimate the integral on a compact support. The limits of the support can be altered by evaluating the basis representations at different discretizations. Providing different equi-spaced mesh grids alters the L_p-metric estimates for the distance between curves. In what

follows, the L_2-metric will be used. Distance estimates are incorporated into R functions (https://github.com/BuscagliaR) that estimate class based on functional KNN (FKNN).

Classifiers are produced in an equivalent fashion to KNN, which was introduced in Section 4.4. For FKNN, the set of training observations that are closest to the test observation, N_0, are now determined by the L_2-metric rather than the Euclidean norm. Estimated probabilities are based on the enumeration of classes within N_0. Supervised FKNN classification algorithms were used to evaluate the SLE plasma thermogram data set. Both the SLE FDO and GCV FDO were investigated on the compact support $T = [45,90]$. Validation of the nonparametric tuning constant was performed during the algorithm with KCV results returned for each requested value.

The results of FKNN classification are given in Table 7 for SLE FDO: the results are remarkably close to the discretized analysis presented in Table 1. Specifically, the mean test set accuracy of original curves is 76.3%, improving to 90.5% for first derivatives. The large jump in mean test set accuracy occurs with significant changes to the mean sensitivity, indicating that first derivative curves improve the discrimination of SLE cases. Specificity is also improved to 88.1%; the combined effects cause the significant jump in classification accuracy for derivative curve, which is nearly equivalent to the results found for discretized data. The second derivative curves return 88.2% mean test set accuracy, slightly higher than classic KNN. Table 8 provides the FKNN results for GCV FDO: in agreement with previous findings, the classification performance of FKNN drops significantly upon using the reduced basis representations. Second derivative GCV FDO curves continue to produce very high sensitivities (94.0%).

NC of SLE plasma thermograms have remarkably reproducible results that are nearly unaffected by the methodology used for estimating distances. Classification using FKNN results in high sensitivity for the unsmoothed SLE FDO. The L_p-metric calculations are flexible to derivative approximations, although at slightly more computational cost than evaluating Euclidean norms. However, the notion of integration on a compact support will provide a useful interrogation tool presented in Chapter 6.

While predictor dimensionality can be altered by reduced sampling from the functional representations, this was shown to have commonly deleterious effects on KNN. The functional alternative is less sensitive to the curse of dimensionality, a concept that will be used to produce FKNN classifiers by altering the compact support being investigated. Changes in the support of the functional representations is equivalent to a change in the limits of integration used in the L_p-metric. The error of the integration problem can be controlled by the sampling from the functional representations. This will produce a family of classifiers that can be combined through ensemble methods, returning predictive models with improved classification performance.

5.5 Combined Functional Covariates

Studied in Chapter 4 was the idea of combining the predictor matrices sampled from the original curves and their first and second derivative approximations. This returned nominally small decreases in classification performance, with first derivative predictors typically outperforming the combined matrices. A similar concept exists for FDA. Instead of a combined predictor matrix with increased dimensionality, the

analogous methodology is to produce models based on the sum of inner products (Ratcliffe et al. 2002). The FLR model can be rewritten as

$$\log\left(\frac{p(X_i)}{1-p(X_i)}\right) = \beta_0 + \sum_k \int_T X_i^k(t)\,\beta^k(t)dt$$

where there are now k possible explanatory functional representations, and each β^k a parameter function. Similar extensions also allow FGSAM and FGKAM to be used with multiple functional covariates.

The results of using multiple functional covariates are shown in Table 9 for SLE FDO. This includes all combinations of the three functional sets: original curves, first derivative, and second derivative curves. Unlike classic LR, in this case dimensionality does not interfere with the computations and results can be obtained from each methodology. The use of multiple functional covariates leads to minor improvements for each FLR classifier. The mean test set accuracies of FGLM improves to up to 77.8% with the use of original curves combined with second derivatives. Additive models also show increases in mean test set accuracy, with FGSAM obtaining 79.3% also for the combination of original with second derivatives.

Kernel methods also produced improved models when evaluating multiple functional covariates. Computational stress of the kernel smoothing and back-fitting routines is evident with the use of FGKAM, with computations taking significantly longer to complete. The results of using multiple functional covariates causes FGKAM to return a mean test set accuracy as high as 79.7% when the first and second derivatives are analyzed together. These models are equally as successful as the FGSAM based on

Original + First Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.748 (0.053)	**0.744** (0.073)	0.753 (0.078)
FGSAM	**0.760** (0.055)	0.738 (0.075)	**0.782** (0.079)
FGKAM	0.734 (0.056)	0.698 (0.085)	0.770 (0.077)
Original + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.778 (0.054)	0.792 (0.076)	0.763 (0.075)
FGSAM	**0.793** (0.051)	**0.798** (0.073)	0.788 (0.074)
FGKAM	0.748 (0.055)	0.696 (0.038)	**0.802** (0.074)
First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.768 (0.056)	0.792 (0.072)	0.743 (0.081)
FGSAM	0.779 (0.051)	0.779 (0.070)	0.780 (0.075)
FGKAM	**0.797** (0.053)	**0.813** (0.072)	**0.781** (0.075)
Original + First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.766 (0.052)	0.773 (0.075)	0.759 (0.075)
FGSAM	**0.776** (0.053)	**0.775** (0.075)	0.777 (0.075)
FGKAM	0.764 (0.054)	0.730 (0.078)	**0.799** (0.076)

Table 9. FLR classification results using multiple functional covariates based on the SLE FDO. Performance is summarized by accuracy, specificity, and sensitivity. Given is the mean and standard deviation for each metric.

original and first derivative functional covariates. The FGKAM model retains high sensitivity to SLE, with small improvements to the specificity when using multiple functional covariates.

An analogous breakdown of GCV FDO is given in Appendix A: Table A3. A typical improvement in derivative based classification performance is observed. The combination of first and second derivative FDOs produces mean test set accuracies of 80.5%, 79.3%, and 74.9% for FGLM, FGSAM, and FGKAM, respectively. This is some of the top performing classification models based on FLR, although these results still fall significantly short of the contemporary penalized LR and FKNN.

5.6 Ensemble Strategies

Models based on derivative curves have shown increased classification performance from nearly all methodologies evaluated. Studied so far have been GLM and FGLM that used combinations of predictors, as well as FGAM analogs. There is no direct method for evaluating multiple FuNC, thus FKNN was omitted from the previous discussion. An alternative method for producing predictive models mixes information from the different derivative classifiers is ensemble learning (Dietterich 2002). Ensemble learning has many applications for classifier combination and can enrich classification performance through improvements of the statistical, computational, and representation problems discussed in Chapter 2.

Let a classifier be any statistical method that estimates class probabilities. Ensemble strategies combine multiple classifiers to produce new estimated predictions. Under a binary classification problem suppose K different classifiers have been produced. Each classifier returns an estimate of the probability, $p_k(X) = P_k(Y = 1|X)$, using $p_k(X)$ to denote the estimated probabilities from the kth classifier. The most fundamental ensemble vote is based on naïve counting of the estimated classes: define the naïve ensemble probability

$$p_E(X) = \frac{1}{K}\sum_{j=1}^{K} I(p_j(X) > \alpha)$$

where $I(p_k(X) > \alpha)$ is the indicator function counting classifiers that produce an estimated probability above the given threshold α, typically set as $\alpha = 0.5$. Naïve ensembles can have difficulties due to ties, and are best employed when ensembles are produced from an odd number of classifiers. For naïve ensembles, ties were broken by random selection of a class.

Weighted ensembles can be produced directly from the estimated probabilities. Define the weighted ensemble probability as

$$p_E(X) = \sum_{j=1}^{K} w_j p_j(X)$$

under the constraint

$$\sum_{j=1}^{K} w_j = 1$$

to ensure properly defined probability estimates. One immediate ensemble prediction is the use of equal weights, returning an analog of the naïve ensemble. Now $p_j(X) \in [0,1]$ reducing the likelihood of ties and gaining more influence on the resulting prediction. Thus, in cases where classifiers suggest opposite class predictions, classifiers that produce higher estimated probabilities will have increased influence on the ensemble probability.

Another potential weighting scheme uses estimated accuracies of the classifiers. Estimated accuracies can be produced from training observations or from resulting KCV test set investigations. Let α_k be the estimated accuracy of the kth classifier, and set

$$w_j = \frac{\alpha_j}{\sum_k \alpha_k}.$$

Accuracy-weighted ensemble probabilities account for the performance of the classifier, increasing the influence of high performing models.

The three ensemble strategies were applied to each classifier used in this study. Based on conclusions that truncated sets only diminished classification performance, only the SLE FULL and GCV FULL sets were considered. The adaptive penalized LR classifiers were also omitted from ensemble studies. Naïve ensembles were produced by mixing original curves, first derivative, and second derivative classifiers. The results of naïve ensembles for SLE FULL are shown in Table 10 and GCV FULL given in Appendix A: Table A4.

Naïve ensembles show positive influence on classification performance, improving nearly all methodologies. LR is slightly improved, while the penalized LR

Naïve Ensemble			
Method	Accuracy	Sensitivity	Specificity
LR	0.726 (0.055)	0.747 (0.086)	0.703 (0.083)
RIDGE	0.912 (0.034)	0.913 (0.049)	0.910 (0.054)
ENET	0.914 (0.038)	0.919 (0.053)	0.909 (0.055)
LASSO	0.906 (0.040)	0.911 (0.055)	0.900 (0.057)
LDA	0.740 (0.056)	0.755 (0.083)	0.725 (0.081)
QDA*	0.898 (0.037)	0.859 (0.059)	**0.937** (0.049)
KNN	**0.919** (0.033)	**0.943** (0.040)	0.893 (0.059)
FGLM	0.737 (0.061)	0.717 (0.082)	0.759 (0.085)
FGSAM	0.764 (0.056)	0.744 (0.079)	0.786 (0.075)
FGKAM	0.761 (0.053)	0.773 (0.076)	0.749 (0.078)
FKNN	0.918 (0.033)	0.933 (0.041)	0.902 (0.053)

Table 10. Performance of naïve ensembles summarized by accuracy, sensitivity, and specificity for all classifiers based on SLE FDO. Naïve ensembles were produced through estimated classes using original curves along with first and second derivatives. Given is the mean and standard deviation for each metric. *Results were obtained using SLE TRUNC predictor set.

classifiers show a general increase in mean test set accuracy. RIDGE is slightly reduced to 91.2% mean test set accuracy, while use of only the first derivative provided 91.6% (Table 1). ENET jumps to 91.4% and LASSO to 90.6% both showing small improvements when using a naïve ensemble. LDA performance is still dramatically low, a consequence of all three curves producing low performance classifiers. QDA was also evaluated, but could only be done so on the TRUNC grid. QDA performance continues to be competitive to the penalized LR methods returning 89.8% mean test set accuracy.

The most significantly altered classifiers are those based on KNN, which produced classifiers with the highest sensitivity to SLE patients. When combined in a naïve ensemble, KNN classifiers produce a mean test set accuracy of 91.9%. The ensemble classifier maintains the property of high mean sensitivity (94.3%) but does so with significant boosts to the specificity of the classifiers (89.3%). This in turn produces one of the highest performing classifiers studied thus far, with equally high performance returning from FKNN. The GCV FDO results are reduced from SLE FDO for RIDGE, ENET, LASSO, and KNN (Table A4). LR, LDA and QDA all show minor increases in performance; these results match earlier studies suggesting that GCV basis representations reduce the performance of penalized LR, while having minor improvements for derivative-based models for discriminant methods.

The success of ensembles can be improved further with the use of weighted strategies. The results of equally-weighted and accuracy-weighted ensembles are shown in Table 11 for SLE FDO, with GCV FDO results given in Appendix A: Table A5. Both equal-weighted and accuracy-weighted strategies produce further improvements to

Equally Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.722 (0.058)	0.738 (0.056)	0.735 (0.054)	0.726 (0.056)
RIDGE	**0.922** (0.034)	0.908 (0.035)	0.904 (0.037)	0.916 (0.034)
ENET	**0.922** (0.034)	0.906 (0.041)	0.903 (0.038)	0.918 (0.036)
LASSO	0.912 (0.036)	0.896 (0.041)	0.899 (0.039)	0.909 (0.039)
LDA	0.741 (0.056)	0.740 (0.056)	0.741 (0.056)	0.741 (0.056)
QDA*	0.895 (0.035)	0.885 (0.039)	0.899 (0.036)	0.897 (0.037)
KNN	0.889 (0.038)	**0.916** (0.035)	0.903 (0.039)	0.923 (0.035)
FGLM	0.741 (0.057)	0.731 (0.058)	0.750 (0.059)	0.748 (0.057)
FGSAM	0.760 (0.053)	0.761 (0.055)	0.768 (0.053)	0.773 (0.054)
FGKAM	0.728 (0.056)	0.745 (0.056)	0.797 (0.051)	0.746 (0.055)
FKNN	0.895 (0.037)	0.913 (0.037)	**0.905** (0.038)	**0.926** (0.034)
Accuracy Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.731 (0.057)	0.738 (0.056)	0.735 (0.054)	0.726 (0.056)
RIDGE	0.922 (0.034)	0.910 (0.034)	**0.905** (0.036)	0.917 (0.033)
ENET	**0.923** (0.034)	0.908 (0.039)	**0.905** (0.036)	0.920 (0.036)
LASSO	0.912 (0.036)	0.897 (0.041)	0.901 (0.038)	0.910 (0.039)
LDA	0.742 (0.057)	0.742 (0.056)	0.742 (0.056)	0.741 (0.056)
QDA*	0.903 (0.032)	0.890 (0.039)	0.902 (0.033)	0.898 (0.037)
KNN	0.899 (0.037)	**0.918** (0.035)	0.902 (0.039)	**0.928** (0.036)
FGLM	0.742 (0.056)	0.733 (0.058)	0.751 (0.058)	0.749 (0.057)
FGSAM	0.761 (0.053)	0.763 (0.055)	0.770 (0.052)	0.775 (0.054)
FGKAM	0.728 (0.056)	0.753 (0.076)	0.805 (0.062)	0.748 (0.053)
FKNN	0.905 (0.035)	0.912 (0.037)	**0.905** (0.038)	0.926 (0.033)

Table 11. Weighted ensemble results for all classifiers based on the SLE FDO. The ensemble probabilities for the combination of all three classifiers (D^0 : original curve, D^1 : first derivative, and D^2 : second derivative) are given. Performance is summarized by accuracy with the test set mean and standard deviation recorded. *Results were obtained using SLE TRUNC predictor set.

classification performance. The equally weighted combination of original curve with first derivative classifiers produces mean test set accuracies of 92.2% for RIDGE and ENET, with LASSO giving 91.2%. Accuracy weighted ensembles return nearly equivalent results of 92.2%, 92.3%, and 91.2%.

KNN classifiers combined from all three curves produce accuracies of 92.3% and 92.8% for equal-weighted and accuracy-weighted ensembles. This suggests that ensembles of NC are competitive with all other methodologies, producing the highest mean test set accuracies of all classifiers. The KNN ensembles are unique in that the inclusion of the second derivative classifiers is important for significant gains in classification performance. This is due to the high true positive rate of the NC. Second derivative KNN classifiers resulted in a mean sensitivity of 93.7%, which work synergistically with the original and first derivative classifiers to produce high accuracy.

The same ensemble strategies were applied to the functional classifiers. The results for the naïve ensemble of original, first derivative, and second derivative classifiers using the SLE FDO basis are given in Table 10. The functional classifiers are not as strongly influenced by ensemble strategies, producing estimates that are lower than evaluating FLR models with multiple covariates. This is likely a consequence of poor performing classifiers, with no information being gained upon mixing individual classifiers. For FGLM and FGAM models, it seems preferable to produce models based on multiple functional covariates. However, such models have significantly lower classification performance than the other methods studied.

The nonparametric FKNN is strongly influenced by ensemble learning. The ensemble of original curves with their first and second derivative classifiers produces mean test set accuracies of 92.6%, nearly equivalent to the contemporary KNN ensemble. The nonparametric classifiers are capable of producing predictive models with the highest classification performance studied. Ensembles of KNN classifiers have been previously studied (Gul et al. 2016), with the suggestion of unique algorithms for identifying the most influential classifiers. The next chapter proposes the ensemble of segmented FuNC. This will be used to build a family of FuNC from each curve or derivative order of interest using segmentation based on altering the limits of integration of the L_p-metric. This family of FuNC can then be subjected to ensemble learning to identify mixtures of classifiers that boost classification performance.

5.7 Conclusions

This dissertation provides a modern in-depth statistical learning analysis of the SLE plasma thermogram. Contemporary methods were studied in Chapter 4 and provided classifiers of high accuracy based on penalized LR. NC were also identified for their high sensitivity to SLE cases. Chapter 5 has now provided an additional statistical evaluation of predictive performance using functional classifiers. FDA was employed along with FLR to produce models based on functional covariates. Alternatives to FGLM, specifically FGAM using spectral smoothing (FGSAM) and kernel smoothing (FGKAM) were also evaluated.

Functional classifiers produced from any individual functional covariate had performance with only minor improvements to contemporary LR. Use of multiple functional covariates improves predictive models, but such classifiers achieve mean test set accuracies of no higher than 80%. FuNC shows significant promise for the classification of SLE plasma thermograms. Similar to contemporary KNN, FKNN produces high sensitivity models with mean test set accuracies as high as 90.5% when using first derivatives. These models results in the highest sensitivity classifiers studied.

Ensemble strategies were then introduced and evaluated using all classifiers discussed in this dissertation. Ensembles improve the performance of all classifiers studied: by combining information from multiple curves classification performance is improved. Ensemble schemes based on naïve voting and weighted prediction probabilities were applied to all classifiers. Penalized LR classifiers, when mixed across multiple derivative orders, return mean test set accuracies as high as 91.9%. This is only a small improvement to overall classification performance, but the improvement occurs for nearly all classifiers studied.

Ensembles of KNN or FKNN classifiers produce large gains to the overall classification performance. Specifically, ensembles of KNN classifiers produce mean test set accuracies as high as 92.8% when combining original curve classifiers with first and second derivative classifiers. A similar increase in classification performance is found for FKNN, which achieves a mean test set accuracy of 92.6% upon the ensemble of all three curve classifiers. This promotes the use of NC techniques for the evaluation of SLE plasma thermograms.

Chapter 6

ENSEMBLE OF SEGMENTED FUNCTIONAL NONPARAMETRIC CLASSIFIERS

6.1 Introduction

Developed herein is the ensemble of segmented functional nonparametric classifiers (ESFuNC) algorithm, which can be applied to multivariate functional data. Classifiers are constructed with data represented as $(X_1, X_2, \ldots, X_p, Y)$, where $X_1, X_2, \ldots, X_p$ are random functional variables defined on a compact interval T and $Y = 0, 1, \ldots, K-1$ the class labels. Supervised classification methods using multivariate functional data have not received significant attention in the literature. Some of the available multivariate functional classification techniques were presented in Chapter 5 (Febrero-Bande and González-Manteiga 2013).

Such methods are of interest even in applications where only univariate functional data are available. Starting with a set of curves, classifiers constructed based on derivative approximations may reduce cross-validation estimates of classification error (Aguilera et al. 2013; Delaigle et al. 2012). Results using derivative SLE plasma thermograms are presented in earlier chapters and agree with these findings. Cross-validation can be used to select the choice of derivative to work with, or information can be combined from original curves and derivatives, as seen in Section 5.6.

The ESFuNC algorithm will be based on FuNC. FKNN classifiers have been suggested as a benchmark for the classification of functional data based on performance and simplicity (Baillo and Cuevas 2008). KNN and FKNN supervised classification has been evaluated in earlier chapters, with both demonstrating strong predictive capabilities for the SLE plasma thermograms. Ensembles of KNN classifiers produced the highest KCV accuracies of all methodologies tested.

FuNC accuracy can be affected if population groups differ significantly only in short subintervals of the compact support. Simulations will be presented to demonstrate such affects, where segmentation of the compact interval into partitions of smaller sub-intervals will improve classification performance. This represents the first segment-wise effort for FuNC, which aims to improve accuracy by selection of the sub-intervals of the functional domain. Similar segment approaches have focused on parametric classifiers and are based on methodologies distinct from those studied here (Delaigle et al. 2012; Li and Yu 2008).

This chapter presents simple techniques that can be employed with virtually any FuNC. For univariate functional data, the approach consists of dividing the original functional domain into non-overlapping segments of equal length. Each segment is considered as a separate functional datum, which is used to produce FuNC. Ensemble combinations of segment subsets are evaluated by LOOCV. Empirical procedures for choosing the number of segments and the optimal subset are investigated. The segments involved in the optimal subset convey interesting information to practitioners, who may

consider the properties of the data generating process that influence the function on the segment intervals.

Multivariate functional data are then examined by evaluating equivalent segmentation and subset approaches on derivative curves. Combination of classifiers based on segments from multiple derivative orders can then be considered. Three strategies are developed for the construction of ensembles based on FuNC. The strategies will be termed greedy, combined, and hierarchical; each of the ensemble strategies are complementary to those introduced in Section 5.6, but take considerations for how segmentation and subset selection may influence how information is mixed across covariates.

The chapter is set up to introduce new methodologies not yet considered in this work, along with computational implementations of the algorithm in Section 6.2. This will be followed by a description of the three ensemble strategies in Section 6.3. Simulations highlighting the impact of the ESFuNC algorithm are presented in Section 6.4, followed by the results of applying the algorithm to the SLE plasma thermograms in Section 6.5. For deeper consideration of the potential empirical improvements the ESFuNC algorithm can offer, benchmark datasets are analyzed in Section 6.6.

6.2 Methodology and Implementation

Let $F = (X(t), Y)$ be an FDO comprised of a set of functional random variables, $X(t)$, defined on the compact interval T with class identifiers, $Y\epsilon\{0, 1, \ldots, K -$

1}. The algorithm is designed to analyze FDOs that have been optimized for basis representation, smoothing penalties, and data reduction techniques (Berrendero et al. 2016; Delaigle et al. 2012; Ferraty and Vieu 2006; Ramsay 2006): each of these steps influences functional representations and may alter classification performance. Considerations as to which derivative orders will be included in the analysis should also be assessed during the production of the primary FDOs. All major R functions developed by the author can be found in the Github repository (www.github.com/BuscagliaR).

6.2.1 Segmentation

The primary method for producing a family of classifiers from univariate functional data will be segmentation of the compact interval T into sub-intervals of equal length $T_1, T_2, \ldots, T_s$. Each sub-interval will be considered as a separate functional datum which will be called a segmented-FDO. Segmentation can be applied to derivatives in addition to the original curves, producing a second method for increasing the number of classifiers considered in the ensemble. $F_{s,j}^{(m)}$ will denote the jth segment of the mth-derivative FDO, partitioned into s total sub-intervals.

This work considers only equal sized partitions of the compact interval. Methods allowing for unequal segmentation patterns are to be considered in future work, and require additional optimization of segment length along with the validation of nonparametric tuning constants. The restriction to equal segment length reduces the computational burden, while providing an easily implemented strategy. As a limit to the

segmentation process, sub-intervals should only be considered such that the length of the interval is larger than the granularity of the original data points. In the analyses that will follow, segmentation was always performed such that interval lengths covered a minimum of three times the granularity of the original mesh.

6.2.2 Functional Nonparametric Classifiers

Introduced in Chapters 4 and 5 was the NC KNN, which returned estimated class probabilities by counting the K-nearest neighbors to the test object. The method is equivalent for either KNN or FKNN, with differences only in the calculation of the distance metric. Considered in the ESFuNC algorithm will be the use of FKNN, with distances measured using the L_2-metric. To broaden the scope of classifiers considered, the algorithm will implement additional NC for estimating class probabilities. KNN can be viewed as an antecedent to the weighted-KNN (WKNN) and Parzen window (PW) classifiers. These methods require a distance metric, but allow for a deeper interrogation of estimated probabilities.

WKNN works equivalently to KNN, but instead of naïve counting of the nearest neighbors, weighting is used based on the distance of the neighbor from the test object. Weights have been historically calculated based on reciprocal distances from the test object (Dudani 1976); modern methodologies produce weights based on a chosen kernel (Hechenbichler and Schliep 2004). Both weighting methods favor neighbors that are closest to the test object, reducing the influence of distant neighbors, with improved

overall classification performance. Ordinary KNN can be considered a special case of WKNN, which corresponds to adapting a uniform kernel.

PW classifiers, also called kernel classifiers, generate estimated probabilities by counting and weighting the neighbors closest to the test object as determined by a given bandwidth (Parzen 1962). Unlike KNN that counts neighbors based on a chosen K, PW classifiers allow for a varying number of neighbors to be selected per test object. Kernel estimators can have significant benefits to supervised learning performance, with improvements to density estimation and feature detection (Muller et al. 2001).

Both NC require the selection of a kernel to produce weights for estimating class probabilities. All calculated distances will be nonnegative, thus asymmetrical kernels are employed. Implemented into the ESFuNC algorithm are the uniform, triangular, and normal asymmetric kernels given below.

$$Uniform\ Kernel: K(d) = I_{[0,1]}(d)$$

$$Triangular\ Kernel: K(d) = 2(1-d)I_{[0,1]}(d)$$

$$Normal\ Kernel: K(d) = \frac{2}{\sqrt{2\pi}} \exp\left\{\frac{-\mathrm{d}^2}{2}\right\} I_{[0,\infty)}(d)$$

Each of the kernels is given in the form used by WKNN, with d representing the calculated distance to the test object. Weighting factors (Δ) are them determined such that $\Delta = K(d)$ for WKNN. The uniform and triangular kernels are defined on the support $[0,1]$, and require distances be transformed to this domain. The normal kernel is defined for all positive distances, but is improved by transforming the minimum observed

distance to the test object to zero. PW classifiers require that each kernel be transformed such that weight factors are determined as

$$\Delta = \frac{1}{h} K\left(\frac{d}{h}\right)$$

with h the bandwidth (Ferraty and Vieu 2006). The PW classifiers only consider neighbors that are within a distance of h to the test object, thus each of the kernels have the support $[0, h]$, including the normal kernel. Any neighbors that fall outside of the bandwidth are given a weight of zero, synonymous with excluding that neighbor from the estimated probability. This creates a dynamic number of neighbors considered for each test case, distinctive from the nearest-neighbor methods.

6.2.3 Stepwise Ensemble

The ESFuNC algorithm constructs ensembles from classifiers produced from segmentation of functional data. To select top performing models, stepwise ensemble strategies have been developed. The two stepwise ensemble methods considered in this work will be termed forward segment selection (FSS) and best segment selection (BSS). Each method combines estimated probabilities from separate classifiers attempting to improve resulting classification accuracy. The stepwise ensembles judge classification performance using LOOCV. Ensembles are created using accuracy-based weights or equal weights, as introduced in Section 5.6. Rank based classifier selection can provide improvements to ensemble classification accuracy (Rokach 2010).

Each segment selection procedure requires that all separate segmented-FDOs be assessed for classification accuracy. Hence, FuNC is performed using each separate segmented-FDO and the LOOCV accuracy of each classifier is computed. FSS is initiated by selecting the segmented-FDO that produces the highest individual LOOCV accuracy. FSS continues by searching through all remaining segmented-FDOs under consideration. The segmented-FDO that returns the largest improvement to LOOCV accuracy when mixed with the highest performing segmented-FDO is retained. FSS continues iterating with each step choosing the segmented-FDO that returns the largest increase in LOOCV accuracy when introduced into the ensemble. FSS terminates when either all segmented-FDOs have been included or when inclusion of any remaining segmented-FDOs does not increase the LOOCV accuracy by a predetermined epsilon.

BSS considers the ensemble of all combinations of segmented-FDOs. BSS is expected to optimize segment combinations and return the mixture of classifiers that produces the highest LOOCV, but does so through a computationally expensive method. Thus some limitations on the number of segmented-FDOs considered by BSS must be made. General rules for modern computing suggests that BSS is computationally feasible when evaluating 25 – 30 segmented-FDOs. Beyond these limits, BSS becomes computationally intractable, with more than a billion combinations having to be considered. BSS is a strong computational tool for evaluating optimized ensemble combinations, but must be used cautiously.

For the above reason, FSS is primarily used when assessing ensemble accuracies. FSS ensembles improve classification accuracy while providing significant

computational advantages (Kohavi and John 1997). Similar FSS procedures have been used to produce models from significantly large libraries with minimal computational cost (Caruana et al. 2004).

6.2.4 Parallelization

The ESFuNC algorithm performs several computationally intensive steps, which can be significantly improved through parallelized computations. Parallel computations were incorporated into the algorithms using the R-packages **doparallel** (Analytics and Weston 2014) and **foreach** (Analytics and Weston 2014b). Distance calculations based on the L_2-metric have low computational cost when considering only a single functional covariate. Segmentation of the functions into s segmented-FDOs increases computational time linearly. Parallel computation of distance metrics for each segmented-FDO is easily implemented and can considerably reduce computational time.

Each of the NC also requires the validation of tuning constants, either neighbor size for WKNN or bandwidth for PW. Parallelization was used to improve the computational time of iterative searches over tuning constants. This is the case especially for PW classifiers, where using a dense grid of bandwidths can be crucial in obtaining an accurate solution. Computational times of BSS were also improved by allowing each combination to be evaluated in parallel. KCV of final ensembles is always conducted, with each fold being evaluated in parallel.

6.3 Multivariate Functional Data Ensemble Strategies

The previous section introduced all of the major elements required to implement the ESFuNC algorithm. Segmentation of a primary FDO redistributes the functional support, producing separate classifiers based on the chosen partition size. Segmented-FDOs can then be used to produce FuNC, where the redistribution of the compact support simply becomes a change in the limits of integration when calculating the L_2-metric. Stepwise strategies for combining the segmented-FDO classifiers then allow for optimal ensemble models to be chosen.

What has not been taken into consideration is how classifiers from multiple derivative orders, or multivariate functional data in general, can be combined. Proposed here are three ensemble schemes that address unique issues in the combination of multivariate functional data. The greedy ensemble strategy (GES) will optimize segmented-FDO ensemble classification for each functional covariate prior to mixing multivariate information. The combined ensemble strategy (CES) will require that ensemble creation be performed simultaneously for all functional covariates, restricting the segment-size to be equal across all covariates. Finally, the hierarchical ensemble strategy (HES) introduces a dependence on the order to which the functional covariates are evaluated. This enforces that critical information determined from earlier functional covariates be retained during the analysis of down-stream covariates. All diagram depictions of the hierarchical strategies were randomly generated and do not reflect SLE classification.

6.3.1 Greedy Ensemble Strategy

The first multivariate ensemble strategy starts from the viewpoint of developing optimized models for each functional covariate. The GES produces optimized segmentation ensembles for each separate covariate; all segmented-FDOs from each functional covariate can then be mixed into multivariate ensembles. A diagram depiction of the GES can be seen in Figure 7; the diagram is generalized to include three FDOs, generically labeled FDO-1, FDO-2, and FDO-3. In what follows, this will correspond to original functional data and their first and second derivatives. However, the method can be used to evaluate distinct functional covariates.

The diagram shows how each individual primary FDO is segmented. The primary FDO is subjected to FuNC, including validation of the nonparametric tuning constant, and the resulting LOOCV accuracy is returned. Segmentation is then increased, demonstrated by the primary FDO being split into two equal sub-intervals. The two segmented-FDOs are each evaluated using FuNC over a grid of tuning constants. This returns estimated class probabilities and LOOCV accuracy estimates for both segmented-FDOs. The ensemble of the two segmented-FDO classifiers is then evaluated by FSS or BSS.

The diagram depicts that for FDO-1, the optimized ensemble at a segmentation size of two includes both segmented-FDOs. On the other hand, FDO-2 is optimized by using only the first half of the functional support. The opposite behavior is given for FDO-3, which uses only the second half of the functional support. FSS or BSS is used at each segmentation size to determine an optimized ensemble providing the highest

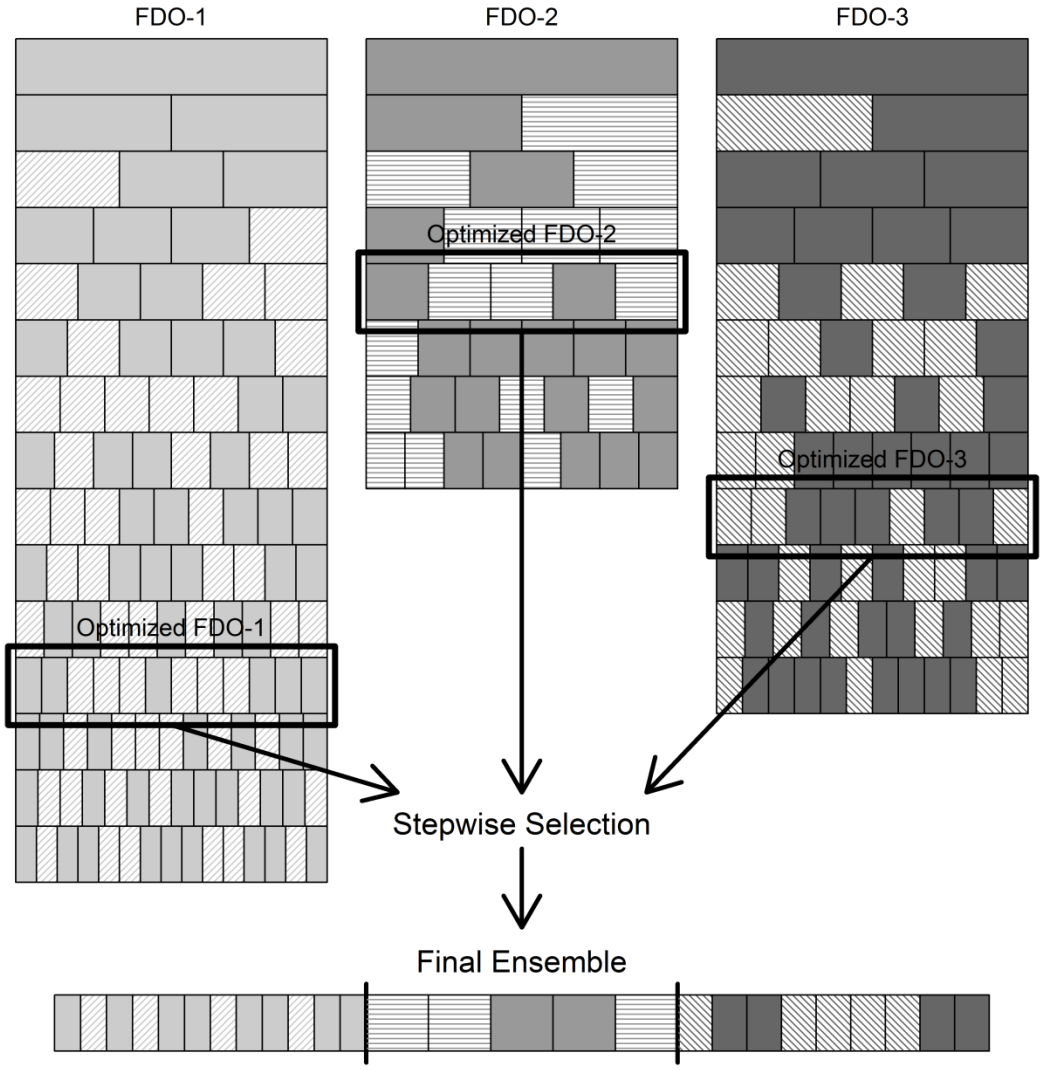

Figure 7. Diagram depiction of the GES for multivariate functional data. The GES determines optimized ensembles from the segmentation of the primary FDO. Once optimized ensembles for each function covariate are determined, all segmented FDOs from each of the different functional covariates are combined using stepwise selection. This can produce a different set of segmented-FDOs included in the final ensemble than was used in the ensemble for any single function covariate. Dark shading indicates the retention of a segmented-FDO in the ensemble model, while light shading indicates the segmented-FDO does not influence the ensemble.

LOOCV accuracy. This also allows for nonparametric tuning constants to be optimized at each segmentation size.

The GES allows segmentation to continue until an optimized ensemble is recovered for each primary FDO. In this case each primary FDOs is subjected to segmentation analysis individually. In the diagram, this corresponds to FDO-1 being partitioned into 12 segmented-FDOs, FDO-2 into 5 segmented-FDOs, and FDO-3 into 9 segmented-FDOs. The diagram depicts that segmentation continues for several additional iterations beyond the optimal ensemble to ensure locally maximized LOOCV accuracy. Importantly, an optimized ensemble classifier is also produced for each FDO. The diagram depicts that of the 12 segmented-FDOs created for FDO-1, only segments 1, 2, 6, 10, 11, and 12 are used in the optimized ensemble. FDO-2 has an optimized ensemble that uses segments 1 and 4, while FDO-3 uses segments 3, 4, 5, 7, and 8.

The segmentation analysis increases classification performance of the separate FDOs as an initial step. This strategy is termed greedy because full optimization of each FDO is done separately with no information shared between functional covariates. The diagram then depicts that all of the segmented-FDOs from each of the three primary FDOs are incorporated into a final stepwise selection process. Although there may be a subset of segmented-FDOs that resulted in an optimized ensemble for the primary FDO, all segments are allowed to enter the final stepwise ensemble search. This allows a wider set of classifiers to be evaluated, and ensures that optimal ensembles will be created based on all information from each primary FDO. The diagram depicts that segments

deemed essential to the final ensemble may change from those essential for the individually optimized FDOs.

The GES is the most computationally expensive of the methods developed, but also gives an in-depth evaluation of each of the primary FDOs. The greedy model essentially produces independently optimized segmentation ensembles, which are then combined in a final step to produce ensembles from multivariate functional data. FSS is typically used during the segmentation analysis of each primary FDO; this is because nonparametric tuning constants are validated at each segmentation size. This requires evaluating a grid of tuning constants at each segmentation size, while additionally evaluating stepwise ensembles for each tuning constant. FSS allows this to be completed with minimal computational cost.

The final stepwise selection is typically done using BSS: based on the diagram, this would suggest that 26 total segmented-FDOs enter the final stepwise selection process. Because tuning constants have all ready been optimized for each FDO, only a single iteration of BSS is required at this stage. The diagram shows that the final ensemble may include distinct segmented-FDOs than those included in the optimized model for each primary FDO. The final ensemble includes segmented-FDOs 1, 3, 5, 6, 8, 9, 11, and 12 from FDO-1, distinct from the segments included when only using information from FDO-1. The final ensemble also includes segmented-FDOs 3 and 4 from FDO-2, and 2, 3, 8, and 9 from FDO-3. This constitutes the final model from the GES. The final ensemble is then subjected to KCV to produce final classification performance metrics, which typically drop slightly from LOOCV estimates.

6.3.2 Combined Ensemble Strategy

The GES required that each primary FDO be optimized prior to any information being shared between the functional covariates. The CES takes a nearly opposite approach, insisting that information be shared across all functional covariates at each stage of segmentation. The CES is summarized in the diagram shown in Figure 8. The combined scheme is initiated by performing FuNC of each primary FDO and using all classifiers during FSS or BSS. The diagram shows that the initial ensemble of FDO-1, FDO-2, and FDO-3 retains all three classifiers.

Unlike the GES, the CES increases the segmentation size of each primary FDO in unison. That is, if segmentation size is increased, all primary FDOs are subjected to the same segmentation. The diagram shows that each of the primary FDOs is then partitioned into two equal sized segmented-FDOs. FSS or BSS is then employed to evaluate an optimized model using information combining all segmented-FDOs. In this case, six segmented-FDOs are mixed, resulting in an optimized ensemble that uses segmented-FDOs 1 and 2 from FDO-1, segmented-FDO 2 from FDO-2, and none of the segmented-FDOs from FDO-3. This process then continues until increasing segmentation size no longer returns ensembles that improve the LOOCV accuracy. The diagram shows an optimized mixture occurs at a segmentation size of 7; this mixture of segmented-FDOs is then taken to be the final ensemble.

Similar to the greedy method, the diagram also depicts that the algorithm will search slightly beyond the optimized segmentation size. This aims to ensure that the ensemble LOOCV has been maximized, and that additional segmentation does not further

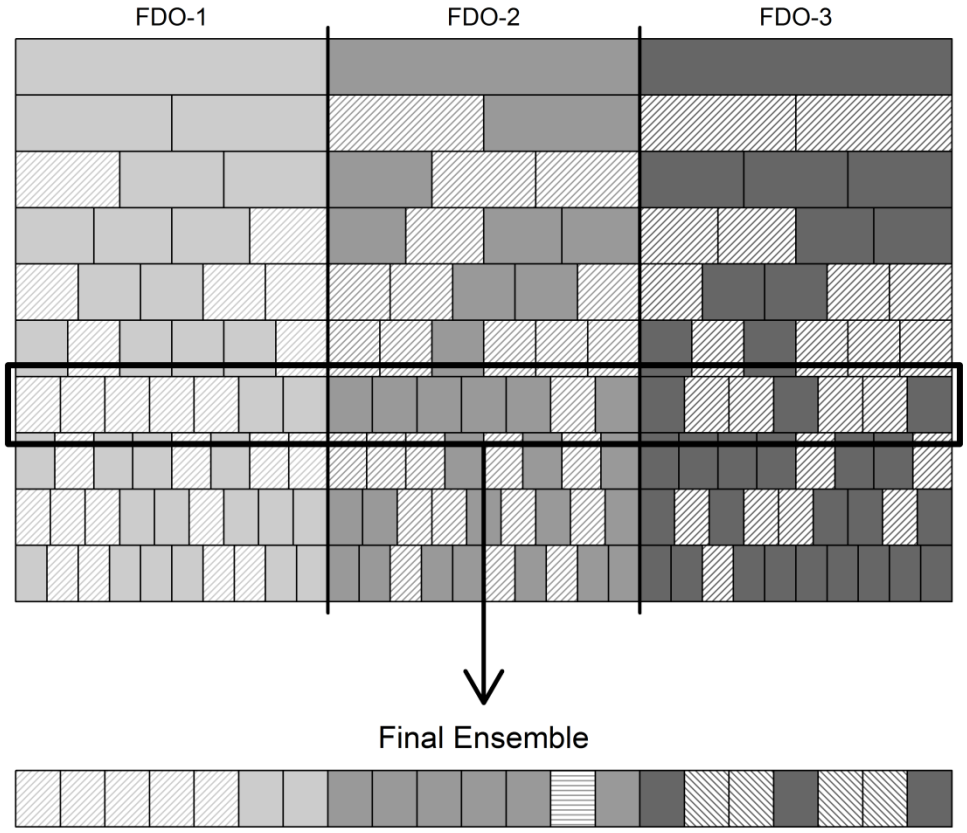

Figure 8. Diagram depiction of the CES for multivariate functional data. The CES enforces equal segmentation sizes for each primary FDO. The CES determines optimized ensembles by mixing all segmented-FDOs from each functional covariate at each segmentation size. Segmentation size is increased until the ensemble produced from mixing all segmented-FDOs it optimized. The final ensemble is taken to be the combination of segmented-FDOs from all primary FDOs that optimizes LOOCV accuracy.

improve classification performance. Because the optimized ensemble is taken to be the final ensemble, the CES is typically evaluated using BSS rather than FSS. Considerations for the tuning constants must also be made. Tuning constants can be computed globally, enforcing all segmented-FDOs to use the same tuning constant, or locally, allowing each segmented-FDO to have a unique tuning constant.

6.3.3 Hierarchical Ensemble Strategy

The first two strategies represent the extremes as to when information from each functional covariate is mixed. The GES only mixed information after producing optimized segmentation ensembles from each primary FDO. The CES enforced that information should be mixed between each primary FDO as segmentation is increased. The HES is a compromise between the greedy and combined methods, allowing information to be shared between primary FDOs while removing the restriction that each FDO must have equivalent segmentation patterns.

A diagram depiction of the HES is given in Figure 9. This scheme is termed hierarchical because of the dependence on the order in which primary FDOs are analyzed. For simplicity, the diagram depicts analyzing in the order FDO-1, FDO-2, and then FDO-3. These are generic labels representing any functional covariate. The diagram shows that optimization of FDO-1 proceeds similar to the greedy ensemble strategy. Classification accuracy is optimized when using only FDO-1 by partitioning into 12 segmented-FDOs. Of these, segments 1, 2, 6, 10, 11, and 12 are essential to the optimized FDO-1 ensemble. To link information between the primary FDOs,

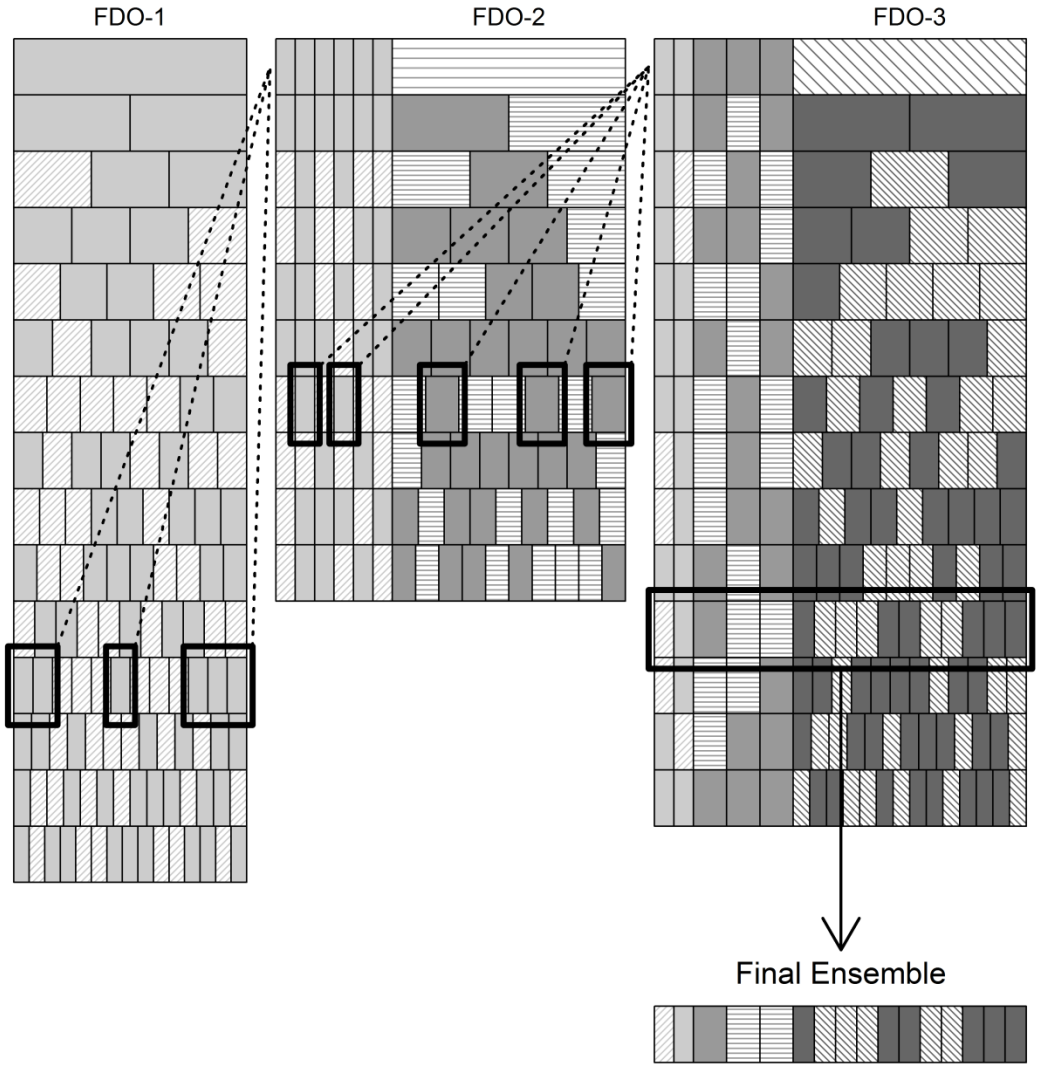

Figure 9. Diagram depiction of the hierarchical ensemble strategy for multivariate functional data. The hierarchical ensemble strategy requires that a choice be made as to the order in which FDOs will be analyzed. The first FDO is then optimized with regards to segmentation, halting when an optimized ensemble of segmented-FDOs has been recovered. The hierarchical strategy is then to retain the segmented-FDOs involved in the optimized ensemble when evaluating the next FDO. This is shown as dashed lines in the diagram connecting FDO-1 to FDO-2. FDO-2 is then optimized in the presence of the retained segmented-FDOs. This allows information to be shared between functional covariates during the optimization process. Ensemble segmented-FDOs are then updated and included during the optimization of FDO-3. The final classifier is based on the optimized ensemble after all FDOs have been analyzed.

optimization of ensemble models involving FDO-2 is evaluated in the presence of the retained segmented-FDOs from optimized FDO-1. This is demonstrated in the diagram by connecting only the segmented-FDOs included in the optimized ensemble of FDO-1 to the optimization of FDO-2. The diagram then shows that inclusion of the primary FDO-2 without segmentation does not influence the results from FDO-1. Optimization through segmentation of FDO-2 then proceeds as usual, but with each stepwise ensemble including the retained segmented-FDOs from FDO-1.

This process continues until a new optimized ensemble combining results from FDO-1 and FDO-2 is achieved. In the diagram, this corresponds to a segmentation size of 7 for FDO-2. The optimized ensemble now uses five segmented-FDOs, two of which were retained from FDO-1, and three of which were found during optimization of FDO-2. The process continues by linking the retained segmented-FDOs to the third primary FDO and optimizing again. This can be continued for any number of functional covariates. The diagram halts at a segmentation size of eleven for FDO-3. The final ensemble is chosen to be the optimized model after iterating through each primary FDO.

Computationally, the HES lies between the greedy and combined strategies. Because the segmented-FDOs retained from each FDO analyzed are typically a truncated set, ensemble sizes do not tend to grow as large as can be found with the GES. This allow for BSS to be implemented in most cases. Validation of tuning constants is conducted at each segmentation size analyzed. When using BSS, smaller tuning constant grids should be used. In practice, deeper searches can be evaluated using FSS, allowing practitioners to search through larger segmentation sizes that are easily handled by FSS.

If smaller grids are to be evaluated, then BSS can typically be employed with some restrictions on how many total segmented-FDOs are allowed to enter the stepwise ensemble algorithms.

6.3.4 Ensemble Strategy Comparisons

Each of the three multivariate functional ensemble strategies has both benefits and disadvantages. The GES allows for an in-depth optimization of each primary FDO, which returns information important to the practitioner related to how well each primary FDO performs when optimized by segmentation. This in turn presents segments of the functional domain deemed important to the final ensemble. Ensembles can also be easily evaluated for all combinations of functional covariates. The greedy method is the most computationally intensive, and typically leads to higher segmentation sizes causing more segmented-FDOs to enter the stepwise ensemble algorithms. This limits the use of BSS for finding optimized ensembles, which can have deleterious effects on the final classifiers.

The CES greatly simplifies the computational complexity while ensuring all information from each primary FDO is included during the construction of ensembles. By enforcing that segmentation patterns be held constant for each primary FDO, segmentation patterns tend to simplify in most cases. Computationally, the number of segmented-FDOs entering the stepwise algorithms grows proportionally to the number of primary FDOs. This allows BSS to be incorporated for small segmentation sizes, but must be switched to FSS as segmentation size grows. Because tuning constants are

validated on a per segmentation basis, the use of BSS is typically restricted to no more than 20 segmented-FDOs. The CES also requires choice of how the tuning constant is validated. Globally versus locally validated tuning constants can influence classification performance and computational time.

The HES represents a more balanced methodology: it allows optimization of each primary FDO to be conducted, but is dependent on a pre-specified sequence in which the FDOs are to be examined. Segmented-FDOs retained in ensemble models are then passed forward as each primary FDO is analyzed. This allows information gained during the optimization of previous FDOs to be passed to later FDOs. By retaining segmented-FDOs during the optimization of additional primary FDOs, unique models that are unlikely to be derived by either the GES or CES are found. The HES is strongly influenced though by the order in which primary FDOs are analyzed: changing the order of analysis can lead to strikingly different segmentation patterns and final ensembles.

6.4 Simulations

This section presents three simulations designed to highlight the importance of segmentation for FuNC and ensembles for boosting classification performance. Considered will be only univariate functional data, without the inclusion of additional functional covariates or derivative curves. The first two simulations evaluate classification of functional data with two classes that differ only on a small interval of the functional support. The difference in the simulations will be the variance of the functional data within the region of difference. The third simulation will create several

small nuisances within the functional data that differ between the two classes. Each simulation will be used to discuss differences between the classification performances of the penalized LR techniques introduced in Chapter 4, FuNC using only the primary curves, and the resulting final ensemble from using the ESFuNC algorithms segmentation optimized ensembles.

6.4.1 Simulation 1

The first simulation analyzes functional data that differ only in a small region of the functional support. The data generating functions for the two classes were set to be

$$F_1(x) = 2e^{-250\,(x-0.25)^2} + 2.25e^{-750\,(x-0.50)^2} + 2e^{-250\,(x-0.75)^2}$$

$$F_2(x) = 2e^{-250\,(x-0.25)^2} + 2e^{-750\,(x-0.50)^2} + 2e^{-250\,(x-0.75)^2}$$

which differ only by 0.25 in amplitude of the peak centered at $x = 0.5$. Random functional variables were generated from $F_1(x)$ and $F_2(x)$ by the addition of normally distributed random noise. That is, random functions $RF_1(x)$ and $RF_2(x)$ are generated by

$$RF_1(x) = F_1(X) + \varepsilon, \;\; \varepsilon \sim N\big(0, \sigma^2(x)\big)$$

$$RF_2(x) = F_2(X) + \varepsilon, \;\; \varepsilon \sim N\big(0, \sigma^2(x)\big)$$

with $\sigma^2(x) = 0.75 - 0.74e^{-260\,(x-0.5)^2}$. This produces a small region near $x = 0.5$ where the two populations are clearly distinct. Random functions were generated using a discretization grid of 200 points over the domain [0,1]; for each class, 200 random

functional variables were generated. This produces a total of 400 samples each with 200 discretized predictors.

Each of the contemporary LR techniques can achieve 100% classification accuracy for this simulation. This is because of the clear region of variance between the classes, which return large coefficients that strongly influence the regression results. It should be emphasized that this simulation was not designed to test the performance of regression methods. Instead, it was designed to show how FuNC performs when the populations differ only at a small region of the functional domain.

Figure 10 presents the random generated functions used in the simulation along with boxplots summarizing the effects of segmentation on classification performance. Classification was performed using WKNN with the uniform kernel (i.e. contemporary FKNN). Functional representations were produced using reduced B-spline basis expansions for minor smoothing. The graph shows vertical lines representing the partitioning of the primary FDO into distinct segmented-FDOs. The fourth segmented-FDO whose domain is $[3/7, 4/7]$ produces a classifier with LOOCV accuracy of 100%.

Boxplots are presented to show how classification accuracy is affected by segmentation. The top performing segment from partition sizes of 1, 3, 5, 7, and 9 were evaluated using KCV. When the full functional domain is used, FuNC is only capable of achieving a mean test set accuracy of 72.2%. This is because most of the differences between the populations can be considered noise, causing distance metrics to be unreliable. As segmentation is increased, classification performance also increases. A segmentation size of 3 returns a mean test set accuracy of 85.9%, which increases to

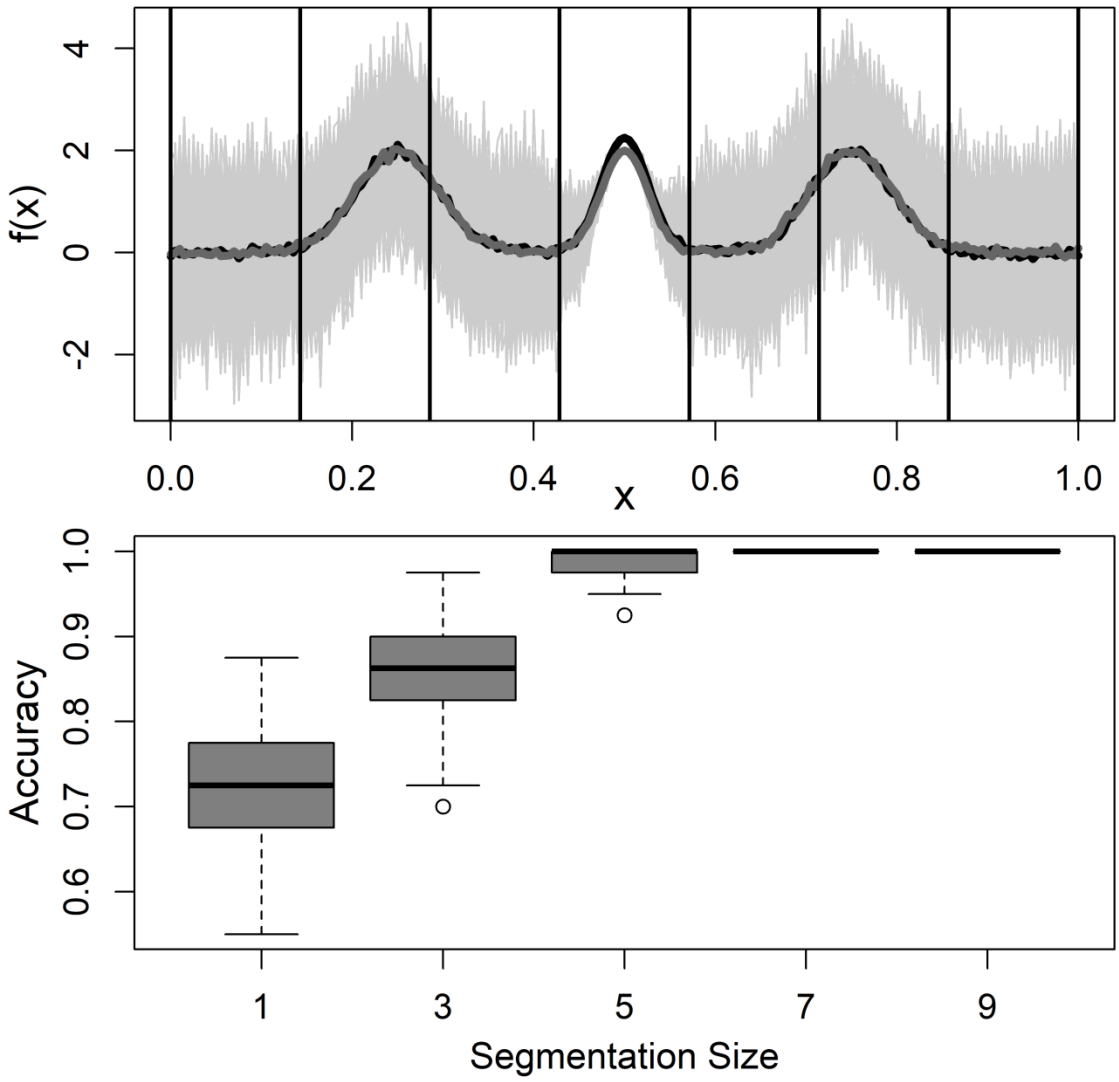

Figure 10. Simulation demonstrating the effect of segmentation on FuNC. The top panel shows the generated random functional variables, with the mean curves for the two classes given as solid lines in black and dark gray. The two population groups are nearly indistinguishable, differing only at a small region of the functional domain centered at $x = 0.5$. Vertical lines represent the partition of the primary FDO into 7 segmented-FDOs. Boxplots then show the change in classification performance as segmentation size increases. At a segmentation size of 7, the central segmented-FDO with domain $[3/7,4/7]$ achieves a mean test set accuracy of 100%.

99.0% for a segmentation size of 5. The ESFuNC algorithm was used to evaluate the dataset and returned an optimal segmentation size of 7. At this segmentation size, the top performing segment produces 100% mean test set accuracy. No changes in KCV accuracy are seen when increased to a segmentation size of 9, indicating that the ESFuNC algorithm identified a segmented-FDO with an optimized classification performance.

This simulation demonstrates a subtle nuance of FuNC. Populations that differ only on a small region of the functional support cannot be adequately classified without truncation of the domain to a more informative interval. The ESFuNC algorithm highlights this, searching for segmentation patterns that can improve classification performance of segmented-FDOs. This simulation does not require ensembles for high performance, but FuNC can achieve 100% mean test set accuracy when the partitioning of the functional support is optimized.

6.4.2 Simulation 2

The second simulation mimics Simulation 1 but uses a constant variance for the normally distributed errors. The same functions, $F_1(x)$ and $F_2(x)$, are used to generate random functions with Gaussian noise, $\varepsilon \sim N(0, \sigma^2)$. Simulation 2 uses a fixed variance $\sigma^2 = 0.75$. The two populations still have a region of significant difference centered at $x = 0.5$, but this region is no longer clearly identifiable. This can be

observed in the graph of the functions given in Figure 11, which shows only the slightest discrepancy in the mean curves for the two populations.

Figure 11 presents a comparison of classification performance for contemporary LR methods against FuNC and the ESFuNC algorithm. Ordinary ML estimation of LR produces a mean test set accuracy of only 56.3%. This can be improved when using penalized LR, matching results found in Chapter 4. LASSO has the highest mean test set accuracy of 65.8%, followed by ENET at 65.5% and RIDGE at 64.1%. This simulation shows how segmentation of functional data improves classification performance. When the entire functional domain is used, FKNN returns a sub-optimal mean test set accuracy of 64.0%, slightly below penalized LR methods.

The ESFuNC algorithm was employed to determine an optimized segmentation pattern and evaluate potential ensembles. The algorithm returns an optimized segmentation size of 7, matching the first simulation. If only the top performing segment is evaluated using KCV, a mean test set accuracy of 69.2% is achieved. This constitutes an improvement of more than 5% in comparison with using the entire functional domain. This also produces improvements over penalized LR, resulting in a more than 3.5% increase in mean test set accuracy. The ESFuNC algorithm additionally determined an ensemble that required more than just the best segmented-FDO. An ensemble of segmented-FDOs 1, 4, and 7 is suggested as optimal based on LOOCV accuracy. When evaluated by KCV, this ensemble has a mean test set accuracy of 69.5%, within the standard error of the best segmented-FDO alone.

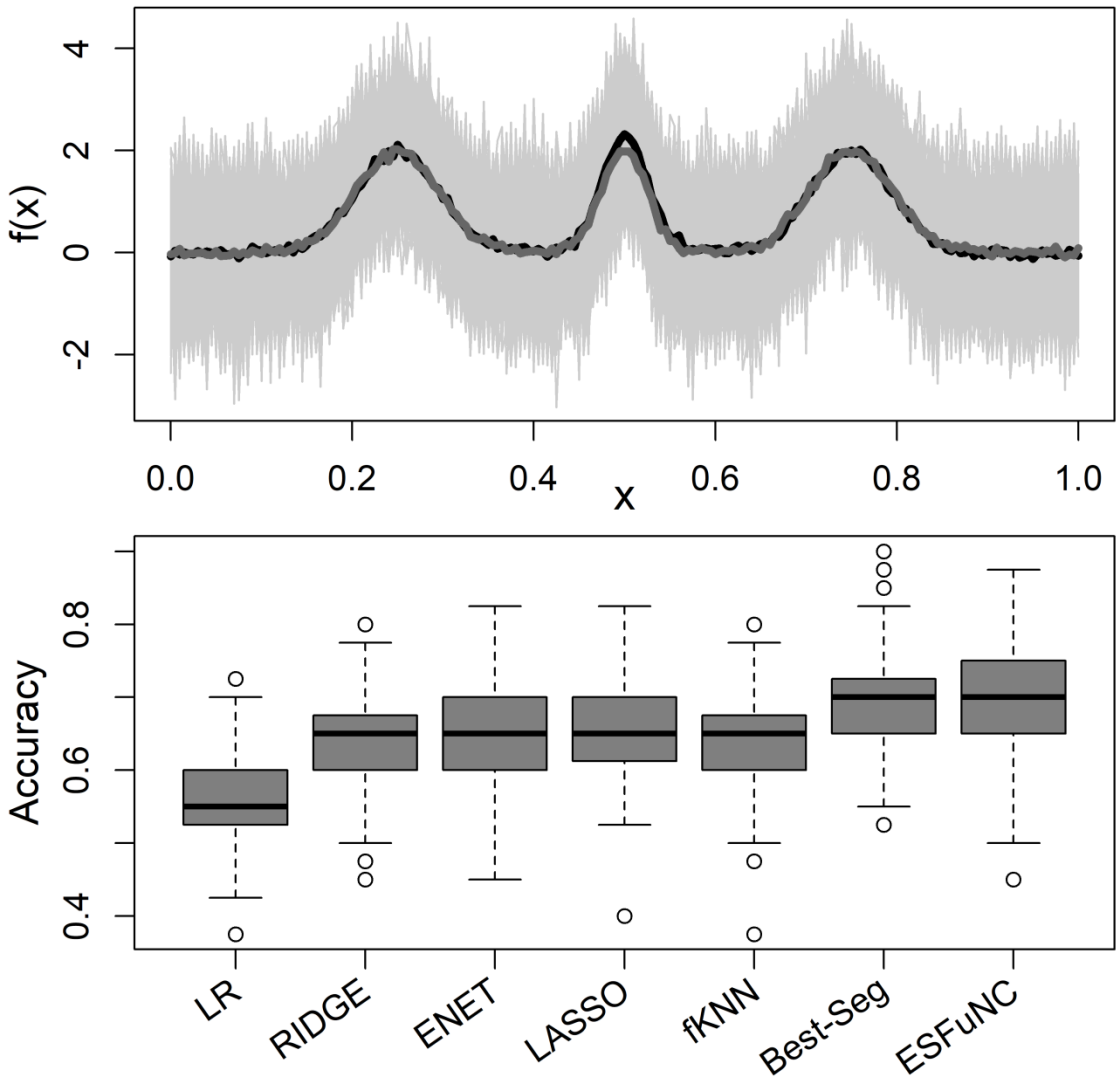

Figure 11. Simulation results involving two populations that differ at one region with equal variance across the entire functional domain. The mean curves for the two population groups have only a small noticeable difference centered at $x = 0.5$. Boxplots show that penalized LR is slightly outperformed by both the best segmented-FDO and the final ensemble produced by the ESFuNC algorithm. In this case, the final ensemble includes two additional segmented-FDOs from the extreme endpoints of the functional domain. This is a consequence of using LOOCV accuracy for decisions on optimized ensembles.

This is a consequence of using LOOCV accuracy for decisions on optimized ensembles. The data generating mechanism that differentiates the two populations should be found only in segmented-FDO 4, which contains $x = 0.5$. However, additional segments included in the ensemble improve LOOCV accuracy slightly. Only a minor gain in classification accuracy is observed, with slight differences in interpretability. If each of the three segmented-FDOs included in the final ensemble are evaluated by KCV separately, it can be quickly determined that segmented-FDO 4 is primarily responsible for the high classification. Segmented-FDO 4 produces a mean test set accuracy of 69.2%, while segmented-FDOs 1 and 7 return significantly reduced results of 51.0% and 53.7%, respectively.

It is evident that the main data generating mechanism for differentiating the population occurs on the interval $[3/7, 4/7]$. The addition of segmented-FDOs 1 and 7, in this case, are erroneous, and cause only a small change in classification performance. These segments are the far extreme points of the functional domain, indicating it may be numerical interference due to the smoothing of the functional representations. Most importantly, both the best segmented-FDO model and the final ESFuNC ensemble outperform penalized LR by more than 3.5% in mean test set accuracy.

6.4.3 Simulation 3

The final simulation introduces three regions of difference between the two population groups. The data generating functions for the two populations groups are now

$$F_1(x) = 2e^{-250\,(x-0.25)^2} + 2.25e^{-750\,(x-0.50)^2} + 2e^{-500\,(x-0.75)^2}$$

$$F_2(x) = 2e^{-250\,(x-0.24)^2} + 2e^{-750\,(x-0.50)^2} + 0.75e^{-1200\,(x-0.74)^2} + 1.5e^{-800\,(x-0.76)^2}.$$

This produces a phase shift in the lower domain of the function, with two equal amplitude peaks being centered at $x = 0.24$ and $x = 0.25$. The central peaks still differ in amplitude and remain centered at $x = 0.5$. Finally, a single peak determines the data generating process near $x = 0.75$ for the first class, whereas two peaks centered at $x = 0.74$ and $x = 0.76$ characterize the data generating process for the second class.

Random functions were generated for each class equivalently to Simulation 2. Errors were introduced as $\varepsilon \sim N(0, \sigma^2)$ with variance fixed at $\sigma^2 = 0.75$. The results of the simulation are summarized in Figure 12. The ESFuNC algorithm returns an optimized segmentation size of 17, which is illustrated by vertical lines in the graph. Along with the segmentation pattern, segmented-FDOs included and excluded from the final ensemble are marked. Darkened segments indicate the segmented-FDO does not influence the final ensemble; segments that are left open are retained in the final ESFuNC ensemble. In this case, the final ensemble includes segmented-FDOs 4, 5, 6, 9, 10, 11, 12, 13, 15, and 17.

The simulation highlights the effectiveness of FuNC over penalized LR techniques. LR is only capable of achieving a mean test set accuracy of 70.3%, similar to findings of Chapter 4, since high collinearity between predictors hinders its performance. This can be recovered by penalized methods, with RIDGE, ENET, and LASSO resulting in mean test set accuracies of 82.1%, 81.9%, and 80.6%, respectively. Each of these

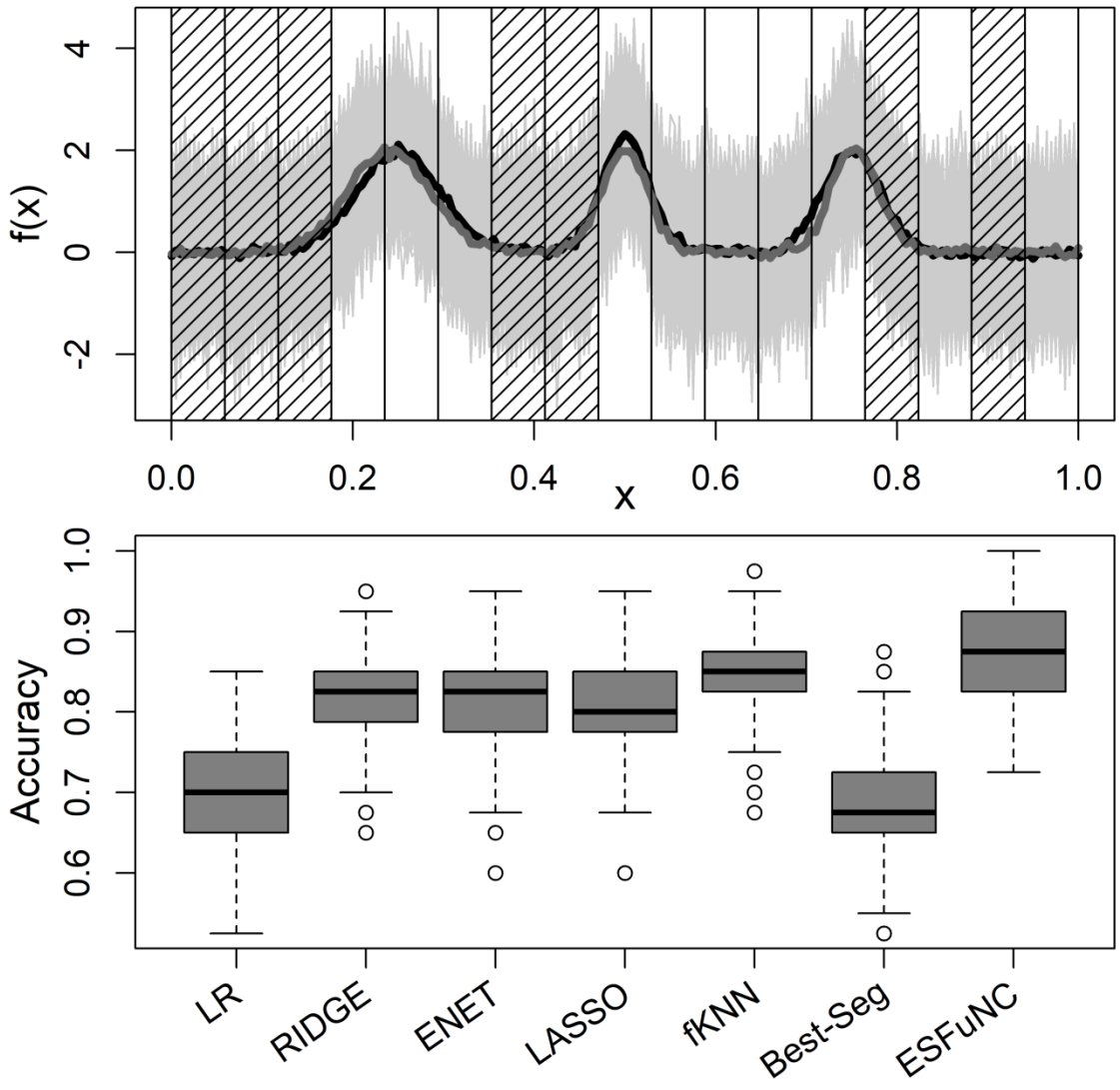

Figure 12. Simulation results involving two populations that differ at three regions with equal variances across the entire functional domain. The mean curves for the two populations have three regions of difference centered at $x = 0.25, 0.5, 0.75$. The graph indicates the segmentation pattern found for the optimized ensemble model. The curve is partitioned into 17 segmented-FDOs. Segments that are not shaded are retained in the final ensemble, while shading indicates the segment does not participate in the final ensemble. Boxplots show that penalized logistic regression is outperformed by both FKNN and the final ensemble determined by the ESFuNC algorithm. When an ensemble of segmented-FDOs is used, classification performance is significantly boosted, although each separate segmented-FDO may produce sub-par performance.

methods are out-performed by FKNN, which achieves a mean test set accuracy of 84.9%. This is an improvement of more than 2.5% from each of the penalized methods, and shows that FKNN is a powerful classification tool when data can be treated by functional representations.

The strength of segmented-FDO ensembles is clearly highlighted by the simulation. The ESFuNC algorithm returns an optimized segmentation size of 17, with the highest performing segmented-FDO being the centralized segment 9. If classification is performed only using segmented-FDO 9, KCV drops to 68.5% mean test set accuracy. This indicates that segmentation no longer improves performance, but instead shows significant dips in accuracy. However, when the ensemble with 9 additional segmented-FDOs is evaluated, classification performance is boosted to a mean test set accuracy of 87.4%. This is a 2.5% increase in performance over FKNN without segmentation, and nearly a 5% improvement over penalized LR.

These three simulations demonstrate the unique advantages of the ESFuNC algorithm, specifically highlighting influences of segmentation and ensemble learning. Simulation 1 shows that segmentation can help distinguish populations that differ only on small regions of the functional support. Simulation 2 provides a similar evaluation, and shows that segmentation of functional data can provide high performance segmented-FDO classifiers. The single segmented-FDO classifier, as well as the final ESFuNC ensemble classifier, both outperform penalized LR methods. Simulation 3 demonstrates the strength of ensemble learning. Although each individual segmented-FDO produces sub-par classification performance, ensembles of segmented-FDOs boost classification

performance. The final ESFuNC ensemble achieves improved classification performance in comparison to both penalized LR and FKNN.

6.5 ESFuNC Analysis of SLE Plasma Thermograms

Classification of the SLE plasma thermogram data set was evaluated using the ESFuNC algorithm. Multivariate functional classifiers were produced by using the original thermogram curves, along with the corresponding first and second derivative approximations. Functional representations incorporated into the analysis correspond to the SLE FDO discussed and used in Chapters 3 – 5, which produces functional representations from unsmoothed B-spline basis expansions. The SLE thermograms were analyzed using each of the three ensemble strategies discussed in Section 6.3.

Specifically, classifiers were based on WKNN and PW: uniform, triangular, and normal kernels were used for WKNN, with uniform kernel being presented as FKNN. PW was incorporated using triangular and normal kernels. Classification performance metrics of each ESFuNC investigation is summarized in Table 12. The optimized segmented-FDOs included in the final ensembles are summarized in Appendix A: Table A6, as well as given graphically in Figure 13.

The GES returns ensemble models with the largest number of segmented-FDOs; this is caused by the independent optimization of each functional covariate, with small improvements occurring as segmentation size increases. This scheme produces high segmentation sizes for each functional covariate, which are then combined in the

Greedy Ensemble Strategy			
Classifier	Accuracy	Sensitivity	Specificity
FKNN	0.936 (0.033)	0.954 (0.035)	0.917 (0.054)
Tri-WKNN	**0.937** (0.030)	**0.962** (0.034)	0.912 (0.055)
Norm-WKNN	0.933 (0.031)	0.944 (0.040)	**0.922** (0.051)
Tri-PW	0.808 (0.051)	0.874 (0.052)	0.741 (0.084)
Norm-PW	0.906 (0.036)	0.937 (0.040)	0.875 (0.063)
Combined Ensemble Strategy			
Classifier	Accuracy	Sensitivity	Specificity
FKNN	0.940 (0.030)	0.951 (0.036)	0.919 (0.052)
Tri-WKNN	0.939 (0.028)	0.940 (0.038)	**0.938** (0.045)
Norm-WKNN	0.940 (0.029)	0.950 (0.039)	0.930 (0.049)
Tri-PW	0.936 (0.030)	**0.952** (0.036)	0.918 (0.051)
Norm-PW	**0.943** (0.029)	0.948 (0.041)	0.937 (0.042)
Hierarchical Ensemble Strategy			
Classifier	Accuracy	Sensitivity	Specificity
FKNN	0.935 (0.030)	0.951 (0.036)	0.919 (0.052)
Tri-WKNN	0.939 (0.032)	**0.956** (0.037)	0.921 (0.052)
Norm-WKNN	0.934 (0.030)	0.943 (0.039)	0.924 (0.051)
Tri-PW	0.935 (0.031)	0.952 (0.037)	0.917 (0.052)
Norm-PW	**0.941** (0.030)	0.955 (0.037)	**0.927** (0.053)

Table 12. ESFuNC results for SLE plasma thermograms for the three multivariate functional ensemble strategies. The table gives the classifier used along with mean and standard deviation of test set performance metrics. Optimized segmented-FDOs are given in Appendix A: Table A6.

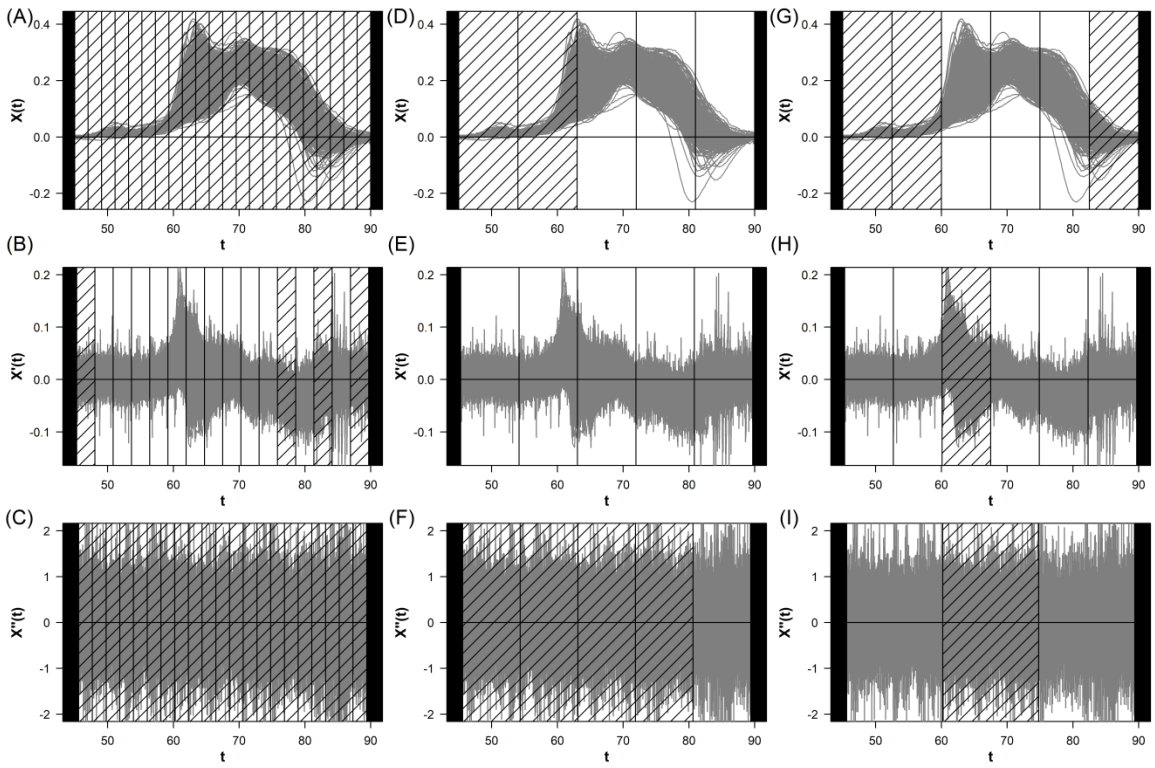

Figure 13. Segmentation patterns of SLE plasma thermograms for top performing ESFuNC ensembles. The original curves along with first and second derivatives are shown for each method. Panels (A) – (C) GES based on triangular-kernel WKNN classifiers. Panels (D) – (F) CES based on normal-kernel PW classifiers. Panels (G) – (I) HES based on normal-kernel PW classifiers. Shaded regions indicate that a segmented-FDO is excluded from the final ensemble.

stepwise ensembles. The GES has significant issues with finding optimized ensembles, requiring the use of FSS for production of final classifiers. If ensembles could be investigated for all potential combinations without computational limitation, it is likely that even higher mean test set accuracies could be accomplished.

The GES returns a maximum mean test set accuracy of 93.7% when based on triangular-kernel WKNN. This classifier has a mean sensitivity of 96.2%, one of the major properties as to why nonparametric methods were considered further. The ensemble uses only information from first derivative approximations, combining 12 segmented-FDOs from a total of 16 potential classifiers. This is a secondary reason why the ensembles may not be completely optimized from the greedy ensemble strategy: BSS ensembles could find optimal combinations of information across all derivative orders, but could not be employed due to a sum of more than 30 total segmented-FDOs being returned from the three functional covariates. FSS produces strong classifiers, with performances higher than found from ensembles of modern methods as evaluated in Chapter 5. This suggests that further tuning of how ensembles are chosen could be an enriching problem for future study.

The GES results also show the difficulties of PW classifiers. Both the triangular- and normal-kernel PW classifiers result in significantly reduced classification performance. This is likely caused by the grid of bandwidth constants evaluated: PW classifiers are sensitive to the bandwidth, and the corresponding mesh may not have optimized values for strong classification performance. This is also a consequence of the large segmentation patterns observed. Smaller segmentation sizes may be covered well

by the bandwidth grid provided; as segmentation size increases, distance metrics will also change, possibly over several orders of magnitude. This may cause the bandwidth grid to miss an optimal value, returning the poor performance observed.

The CES is the highest performing of all the methods tested. Mean test set accuracies as high as 94.3% are found when using normal-kernel PW classifiers. There is remarkable similarity between all classifiers used, with segmentation sizes of 5 being returned for 4 of the 5 classifiers tested (Table A6). The segmentation patterns are also highly simplified relative to the GES. This mostly results from the limited increases in segmentation size. However, the CES finds stronger ensembles with lower segmentation patterns. This suggests that the greedy ensembles are either over-fitting segmentation sizes, or that higher performance models are possible but were not identified by FSS. The highest performing classifier also has significantly high specificity and sensitivity, unlike the solely high sensitivity models observed from KNN and FKNN in Chapter 5. Mean test set sensitivity of 94.8% is achieved with a mean specificity of 93.7%, representing a model with the highest combination of sensitivity and specificity found in this study.

The HES was implemented starting from the original curves followed by first and then second derivatives; this returns maximized mean test set accuracy when based on normal-kernel PW classifiers. The mean test set accuracy of 94.1% is only minimally reduced from the CES, and is achieved with segmented-FDOs distinct from those of the combined scheme. The optimized combination provides a higher mean sensitivity of 95.5%, but with reduced specificity of 92.7%. The ESFuNC models achieve an overall

balance between sensitivity and specificity for SLE classification, with the most balanced ensembles returned from the CES.

The results show that the ESFuNC algorithm is capable of achieving improved classification performance over contemporary methods and ensembles using the full functional domain. Segmentation increases the number of classifiers considered; these segmented-FDOs are then combined using stepwise ensemble algorithms, producing models with high accuracies, driven by balanced specificity for SLE and sensitivity to non-SLE alternatives. The mean test set accuracy of 94.3% achieved from the combined scheme represents a classification performance improvement of over 5% relative to the most recently published SLE plasma thermogram investigations (Garbett et al. 2017). It is worth emphasizing that this recent study was able to achieve 89% mean test set accuracy, but required information from multiple sources. The ESFuNC algorithm represents a novel approach to building high performance ensembles using only information from plasma thermogram.

6.6 ESFuNC Analysis of Benchmark Data

The ESFuNC algorithm was additionally applied to two benchmark data sets to further evaluate the overall performance of the proposed technique. The first benchmark set, henceforth named the Tecator set, comprises near infrared transmission spectra collected for 215 finely chopped meat samples. The data was collected over a wavelength range of 850 – 1050 nm using a Tecator Infratec Food and Feed Analyzer, with fat content reported for each sample (Thodberg 1995). Observations were

partitioned into 138 low and 77 high fat content samples based on a cutoff of 20%. FDA was applied using unsmoothed B-spline representations. The classification performance of the ESFuNC algorithm, using each multivariate ensemble strategy, is summarized in Table 13. Final ensemble optimized segmented-FDOs are summarized in Appendix A: Table A8.

The ESFuNC algorithm performs exceptionally well for the Tecator data. This benchmark data highlights that the ESFuNC algorithm is capable of capturing simple specifications that lead to classifiers of high accuracy. The GES returns a mean test set accuracy of 99.8% when using the triangular-kernel WKNN classifier. The final ensemble utilizes three segmented-FDOs combined from first and second derivative classifiers, boosting performance beyond modern literature investigations. Of significant importance is the regions returned by the ESFuNC algorithm. The suggested important region for first derivative classifiers is the 850 – 950 nm range, while second derivative classifiers are based on the wavelength ranges 850 – 883 nm and 917 – 950 nm. A graphical representation of the top performing ESFuNC segmentation pattern is given in Figure 14.

Dissimilarity representations based on FDA provided improvements over performing classification on the functional data alone (Porro-Muñoz et al. 2011). The DR-FDA procedure produced test accuracies as high as 99.5%, and focused primarily on feature selection from the original curves. Supervised classification combined with support vector machines was shown to produce a set of 6 unique clusters. Highlighted in the cluster analysis is the selection of a cluster near the 930 nm range, in agreement with

Greedy Ensemble Strategy			
Classifier	Accuracy	Sensitivity	Specificity
KNN	0.995 (0.015)	0.987 (0.040)	**1.000** (0.005)
Tri-WKNN	**0.998** (0.011)	**0.994** (0.031)	**1.000** (0.000)
Norm-WKNN	0.997 (0.011)	**0.994** (0.030)	**1.000** (0.005)
Tri-PW	0.907 (0.060)	0.759 (0.157)	0.990 (0.028)
Norm-PW	0.964 (0.037)	0.899 (0.103)	**1.000** (0.000)
Combined Ensemble Strategy			
Classifier	Accuracy	Sensitivity	Specificity
KNN	**0.995** (0.015)	0.987 (0.040)	0.999 (0.007)
Tri-WKNN	0.992 (0.018)	0.979 (0.049)	0.999 (0.007)
Norm-WKNN	0.993 (0.017)	0.980 (0.048)	**1.000** (0.000)
Tri-PW	0.994 (0.016)	0.985 (0.043)	0.999 (0.009)
Norm-PW	0.994 (0.016)	**0.994** (0.027)	0.993 (0.021)
Hierarchical Ensemble Strategy			
Classifier	Accuracy	Sensitivity	Specificity
KNN	**0.995** (0.015)	0.987 (0.040)	**1.000** (0.005)
Tri-WKNN	**0.995** (0.015)	0.987 (0.040)	**1.000** (0.005)
Norm-WKNN	**0.995** (0.015)	0.987 (0.040)	**1.000** (0.005)
Tri-PW	0.990 (0.020)	0.974 (0.056)	**1.000** (0.005)
Norm-PW	0.994 (0.016)	**0.996** (0.027)	0.993 (0.021)

Table 13. ESFuNC results for the Tecator data set for all three multivariate functional ensemble strategies. The table gives the classifier used along with mean and standard deviation of test set performance metrics. Optimized segmented-FDOs are given in Appendix A: Table A7.

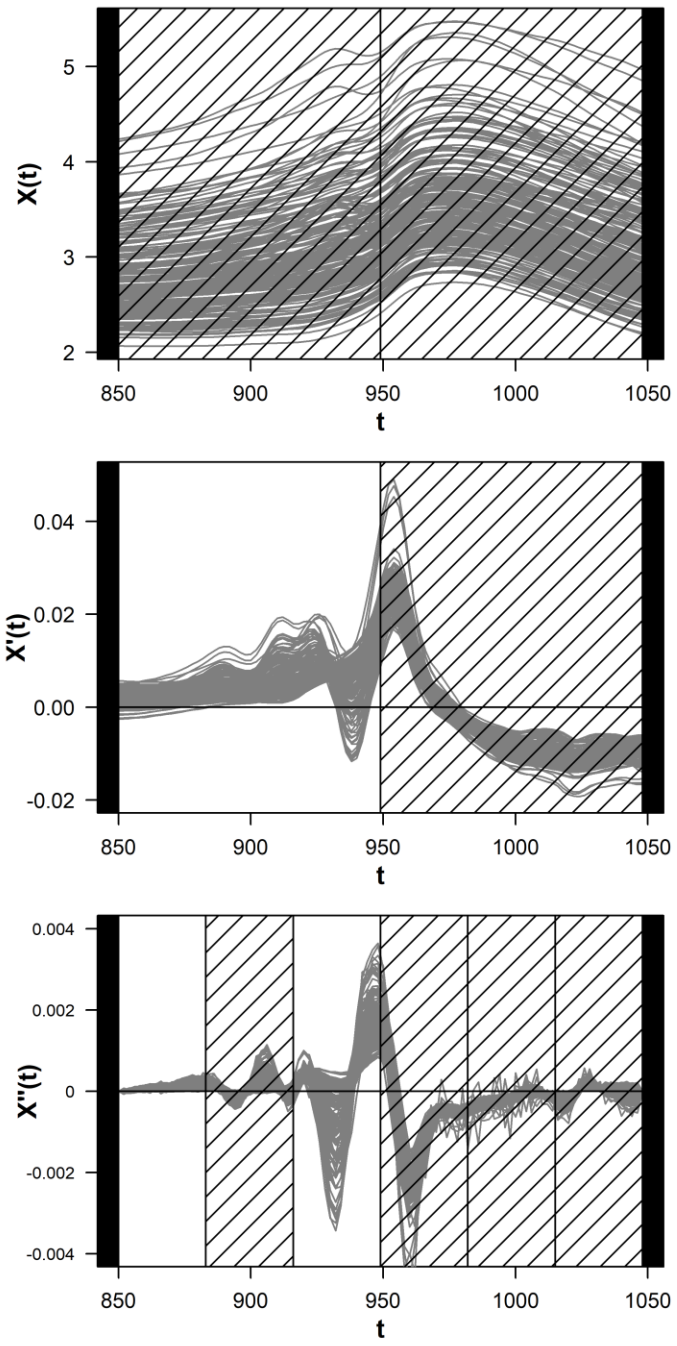

Figure 14. ESFuNC segmented-FDO pattern for the Tecator data using the GES with triangular-kernel WKNN. Shaded regions indicate segmented-FDOs which are omitted from the final ensemble. The figure shows that the upper half of the functional domain of first derivative curves is used in combination with two segmented-FDOs from the second derivative.

the algorithm developed here. The supervised clustering SVM analysis produced average test accuracies of 97.7%, and includes significant dimensionality reduction (Krier et al. 2009).

Functional segment discriminant analysis (FDSA) has also been applied to the Tecator data set (Li and Yu 2008). This method has a similar philosophy to the procedure developed here and proposes a combination of linear discriminant analysis and support vector machines. LDA is used to produce feature selection, which is then implemented into SVMs. Discriminant analysis results in segmentation of the underlying data curves comparable to the procedure developed in this work. FDSA produces test error rates of 97.1% and 98.9% based on the analysis of first and second derivatives, respectively. Wavelengths of 905, 935, and 1045 nm were found to have the highest frequency of inclusion in curve segments. The low wavelengths agree well with the support range of 850 – 950 nm found from the first half of the functional support, and in addition agree with the 917 – 950 nm range derived from second derivative classifiers. The ensemble procedure suggests that FDSA may be further improved through a combination of the results found from the first and second derivatives.

ESFuNC analysis of the Tecator benchmark data set returns a mean test set accuracy of 99.8%, which is higher than what has been reported in published literature so far. Importantly, the segmented-FDOs found in the optimized ensembles have excellent agreement with other techniques used to identify critical regions. This supports that the ESFuNC algorithm can both improve empirical classification results while offering practitioners useful insights into regions critical to discriminating populations.

A second benchmark data set was also investigated using the ESFuNC algorithm. The Phoneme data set consists of 2000 recorded speech frames divided into 5 distinct classes. The set comprises 400 scans from five unique phonemes ("sh", "dcl", "iy", "aa", "ao"), each analyzed using 150 point log-periodograms taken from the TIMIT database (Friedman et al. 2001; Hastie et al. 1995). FDA was applied using a reduced B-spline basis expansion consisting of 50 basis functions for minor smoothing of the speech frames with no roughness penalty. The result of the ESFuNC algorithm for all classifiers and ensemble strategies is summarized in Table 14. The optimized segmented-FDOs for each final Phoneme classifier are summarized in Appendix A: Table A8

The ESFuNC algorithm returns nearly equivalent models for all classifiers and multivariate ensemble strategies investigated. Nearly all final ensembles incorporate segmented-FDOs from original curves and both first and second derivatives (Table A8). The hierarchical ensemble strategy using the normal-kernel PW classifier produces a mean test set accuracy of 93.5%. Sensitivities are reported for the phoneme "dcl", with all models producing mean sensitivities of 100%. In contrast, mean specificities drop to nominally 91% due to the difficulty in discriminating between the "aa" and "ao" phonetic speech frames.

Phoneme classification is well studied in the literature, with the TIMIT database having been used in a wide variety of studies. Previous investigations based on deep neural networks and KNN classifiers demonstrated that choice of distance metric could influence phoneme speech frame classification accuracy (Rizwan and Anderson 2014). However, the KNN classifiers and distance metrics were based on point-wise

Greedy Ensemble Strategy			
Method	Accuracy	Sensitivity	Specificity
KNN	0.928 (0.016)	**1.000** (0.002)	0.910 (0.020)
Tri-WKNN	0.925 (0.016)	**1.000** (0.000)	0.906 (0.020)
Norm-WKNN	**0.930** (0.016)	**1.000** (0.003)	**0.912** (0.020)
Tri-PW	0.924 (0.016)	0.997 (0.008)	0.906 (0.020)
Norm-PW	0.927 (0.016)	0.998 (0.007)	0.909 (0.019)
Combined Ensemble Strategy			
Method	Accuracy	Sensitivity	Specificity
KNN	0.929 (0.016)	**1.000** (0.002)	0.911 (0.019)
Tri-WKNN	0.929 (0.017)	0.998 (0.008)	0.912 (0.020)
Norm-WKNN	0.930 (0.015)	0.999 (0.004)	0.913 (0.019)
Tri-PW	0.929 (0.016)	0.999 (0.004)	0.911 (0.020)
Norm-PW	**0.932** (0.016)	**1.000** (0.000)	**0.915** (0.020)
Hierarchical Ensemble Strategy			
Method	Accuracy	Sensitivity	Specificity
KNN	0.932 (0.015)	**1.000** (0.000)	0.915 (0.019)
Tri-WKNN	0.932 (0.014)	**1.000** (0.000)	0.915 (0.018)
Norm-WKNN	0.931 (0.016)	0.998 (0.007)	0.915 (0.020)
Tri-PW	0.931 (0.016)	0.997 (0.008)	0.914 (0.020)
Norm-PW	**0.935** (0.015)	**1.000** (0.002)	**0.919** (0.019)

Table 14. ESFuNC results for the Phoneme data set for all three multivariate functional ensemble strategies. The table gives the classifier used along with mean and standard deviation of test set performance metrics. Optimized segmented-FDOs are given in Appendix A: Table A8.

calculations, and accuracies were only as high as 78%. Hence, FuNC may provide significant improvements to the classification of phonetic patterns. Critical to the high test set accuracy achieved by ESFuNC is that the optimal ensemble combines segments across all three curves. Such an approach has not been implemented in previous analyses of the Phoneme data set.

The FSDA approach by (Li and Yu 2008) achieved a mean test accuracy of 82.5%, significantly lower than the nominally 92.5 – 93.5% mean test set accuracies attained here. The results presented here agree well with the original investigations reported in (Hastie et al. 1995) using penalized discriminant analysis, which produced mean test set accuracies of 92.5%. Thus, the proposed ESFuNC algorithm produces high accuracy classifiers for the Phoneme data set, returning equivalent or slightly improved models from previously published work.

This application demonstrates the importance of combining segments from multiple derivatives. Second derivative classifiers provide suboptimal classification performance, producing mean test set accuracies of nominally 60% or lower, when used alone. However, when combined with higher performing segmented-FDOs, an overall improvement to the classification of the phonetic speech frames is observed. This is believed to be one of the important advantages of the ESFuNC algorithm, which allows for multiple functional covariates to be combined. Such methods could improve a variety of studies, with the methodologies developed in this work having simple extensions to previous literature investigations.

6.7 Conclusions

The ESFuNC algorithm has been developed and shown to provide empirical improvements to classification of functional data. The algorithm utilizes segmentation to produce an increased set of classifiers that are combined to form an ensemble using stepwise algorithms. WKNN and PW are incorporated into the algorithm, along with uniform, triangular, and normal kernels. Ensembles are produced from FSS and BSS, allowing for either improved computational speed or intensive investigation of all possible classifier combinations, respectively. Developed in this work are three multivariate functional ensemble strategies. Each strategy has advantages and disadvantages, with differences in how classifier information is mixed across multiple functional covariates.

Several key aspects of the ESFuNC algorithm were explored by simulation. Simulations demonstrate how FuNC can be inhibited when population groups differ only on small regions of the functional domain. The segmentation approach implemented within the ESFuNC algorithm uses truncated ranges of the functional domain, termed segmented-FDOs, to produce improvements in FKNN performance. Further simulations showed empirical improvements of the ESFuNC algorithm over LR techniques. The ESFuNC algorithm is capable of improving overall classification performance in a variety of scenarios. The algorithm achieves improvements either through selection of segmented-FDOs that improve overall classification, or by ensembles of segmented-FDOs which boost performance when used in tandem.

The ESFuNC algorithm was applied to the SLE plasma thermogram data set displaying mean test set accuracies as high as 94.3%, an improvement of more than 5% over recent literature investigations. Minor differences were observed between the three multivariate ensemble strategies, with the CES and HES returning the most simplified and interpretable models. In addition to supplying high accuracy classifiers, the ESFuNC algorithm balances sensitivity to SLE patients with specificity for non-SLE alternatives. NC were determined to give high sensitivity models from the investigations presented in Chapters 4 and 5. This property was utilized in the ESFuNC algorithm, and upon ensemble of several segmented-FDO classifiers, both high sensitivity and high specificity models were produced. The ESFuNC ensembles offer significant improvements for the identification of SLE using only plasma thermogram data. Such models make plasma thermograms a promising diagnostic technique for SLE identification, with potential applications to other autoimmune diseases.

Benchmark data sets, Tecator and Phoneme, were used to broaden the scope of potential uses of the ESFuNC algorithm. Application of the ESFuNC algorithm to the Tecator data set results in mean test set accuracies as high as 99.8%. The functional domain intervals suggested by the segmented-FDOs chosen from the ESFuNC algorithm agree with previous literature investigations. This demonstrates that the ESFuNC algorithm can both provide important empirical improvements to the classification of functional data, but also provides insights about which regions of the functional domain are critical for the discrimination of populations.

Phoneme results demonstrate that the ESFuNC algorithm can also be utilized for classification when more than two populations are present within the data set. Final ensembles produced mean test set accuracies as high as 93.5%. Importantly, the phoneme results show the strength of combining segmented-FDOs generated from multiple functional covariates. The classification performance is contributed to a mixing of information across multiple derivative orders; such mixing was not been utilized in previous studies. It could be pertinent to evaluate discrimination of the "aa" and "ao" phonetic speech frames, as these two populations are responsible for nearly all misclassifications. ESFuNC models built to discriminate these two speech frames specifically have yet to be studied.

The ESFuNC algorithm has significant potential for empirical improvements to classification. The algorithm is designed to boost performance through segmentation and ensemble strategies. The final ensemble classifiers also provide unique insights into regions of the functional support that may be critical for discriminating between populations. Significant updates to the algorithm, changes to the computational implementations, and ensemble estimation beyond stepwise methods are excellent sources for future research.

Chapter 7

FPCA-BASED CLASSIFIERS AND PER-PIXEL ENSEMBLES

7.1 Introduction

The supervised classification algorithms developed in Chapter 4 are extended to predictors based on FPCA. A learning algorithm is created to train classifiers using different sets of FPCs, whose scores are used as predictors. The number of FPCs impacts FPCA-based classifier performance; learning algorithms return optimized classification accuracies by iterating through an increasing set of FPCs. Higher accuracy classifiers are produced for SLE plasma thermograms when based on FPC scores in comparison to using primary data predictors.

The concept of segmentation introduced by the ESFuNC analysis suggests that ensembles may boost classification performance. This chapter explores how classifiers based on a single predictor (pixel) can be combined to produce effective ensembles. Several ensemble building strategies using per-pixel classifiers (PPC) are explored. Per-pixel ensembles (PPE) approach the effectiveness of the ESFuNC algorithm when using penalized LR to estimate ensemble weights. Penalized LR based on the fused-LASSO penalty (Tibshirani et al. 2005) is introduced as a method capable of grouping pixels while taking neighboring pixel association into account. Fused-LASSO returns unique

information regarding how pixels may be grouped while simultaneously allowing for estimation of LR classifiers.

7.2 Supervised Classification using FPC Scores

Chapter 4 illustrated that several of the classifiers tested were inhibited by multicollinearity. Estimation of LR models using PCA can improve predictive performance with collinear predictors (Aguilera et al. 2006). PCA-based LR was implemented previously for classification of SLE plasma thermograms (Garbett and Brock 2016); however, the authors chose only to investigate classifiers based on the first six components, resulting in sub-par performance of only 71% test set accuracy. This section reevaluates the performance of contemporary classifiers using FPCA-based predictors.

FPCA was introduced in Chapter 3 to evaluate modes of variability within the SLE plasma thermogram data set and its derivative approximations. FPCA is analogous to PCA, allowing for variability to be explained within a functional data context. Introduced was the computation of FPC scores, which provide the contribution of each functional covariate to each principal component. FPC scores were implemented into the supervised learning algorithms developed in Chapter 4. FPC scores are used as predictors having the added benefit that scores from different components do not suffer from multicollinearity.

The algorithms were updated to allow for supervised learning to occur across a grid of FPCs: algorithms were designed to perform KCV using predictors based on the first k FPCs, with $k = 2, 4, \ldots, 350$. Each of the classifiers introduced in Chapter 4 were evaluated; the supervised learning results based on ML estimation (LR) and LASSO are presented in Figure 15. The figure summarizes KCV accuracy for classification of SLE plasma thermograms as boxplots at each value of k (FPCs), with performance shown for FPC scores derived from the original curves along with first and second derivatives.

It is evident from the supervised learning output that the classification performance depends strongly on the number of FPCs included. When too few FPCs are used as predictors, performance suffers, returning KCV accuracies of 70 – 75%; these values agree with the results of (Garbett and Brock 2016). Figure 15 highlights the improvements to classification performance as the number of FPCs is increased. The supervised learning algorithm is able to identify the number of FPCs to include as predictors to optimize test set accuracy. A summary of the KCV results for all seven classifiers is given in Table 15.

Figure 15 shows that both LR and LASSO have significant improvements to classification performance as the number of FPCs included in classifier estimation is increased. LR performance increases sharply until 48 FPCs are included, with a region of 46 – 60 FPCs giving high mean test set accuracies for original curves along with first and second derivatives. A mean test set accuracy of 91.6% can be achieved based on LR when using 48 FPCs for the original curves. This is accompanied by a mean test set

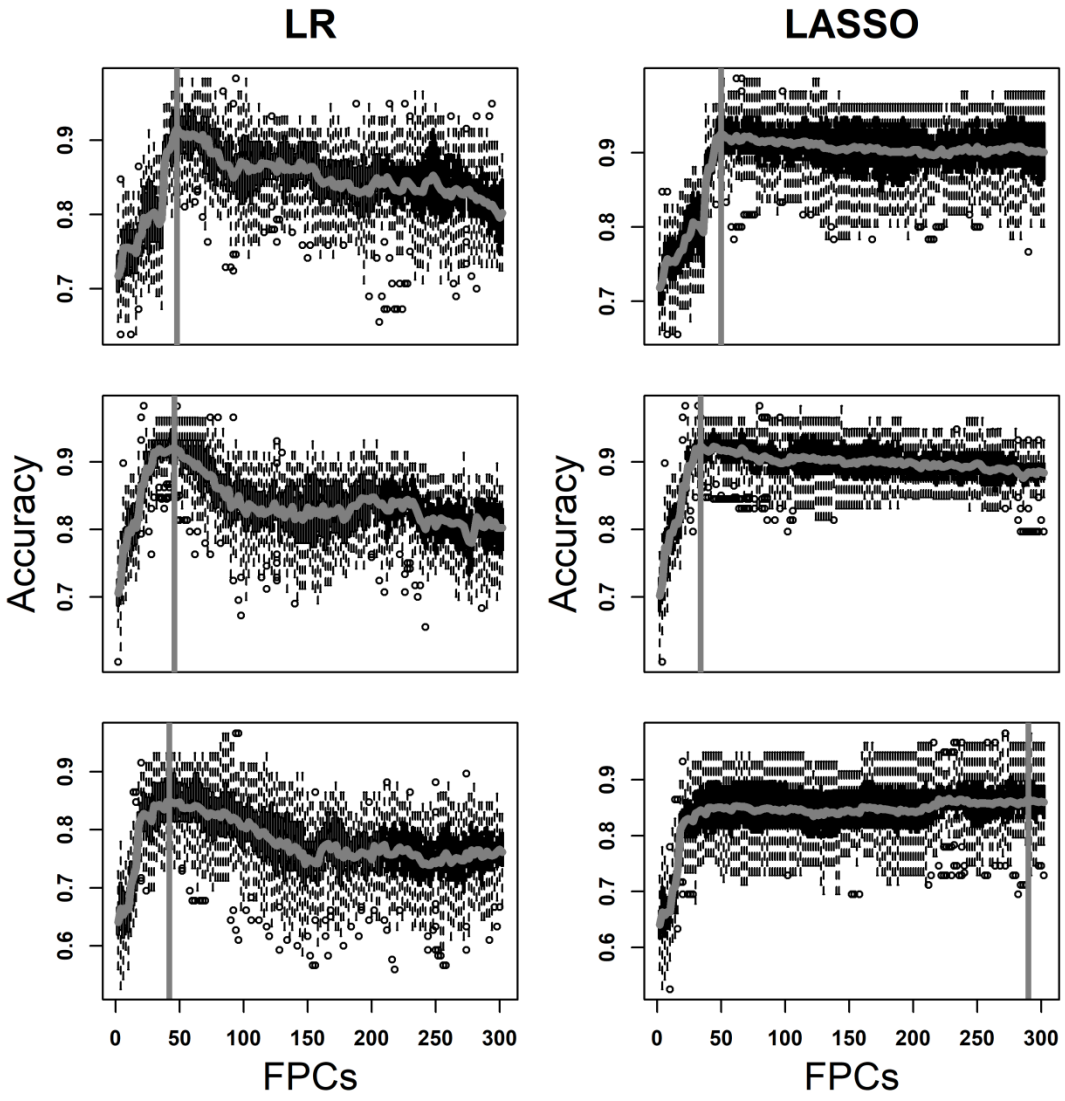

Figure 15. Summary of FPCA-based LR and LASSO classification performance using an increasing number of FPCs. Results are shown based on FPCA of the SLE FDO original curves and its first and second derivatives. Original curves are shown on top, followed by first derivatives in the middle and second derivatives on the bottom. Boxplots of classification accuracies resulting from KCV are given for each value of k. Grey line indicates the mean test set accuracy at each FPC. The vertical grey line represents the number of FPCs that returns the highest mean test set accuracy.

Original Curves				
Method	FPCs	Accuracy	Sensitivity	Specificity
LR	48	0.916 (0.037)	0.911 (0.041)	0.921 (0.056)
RIDGE	50	**0.935** (0.038)	0.928 (0.037)	**0.942** (0.057)
ENET	50	0.932 (0.030)	0.926 (0.039)	0.938 (0.051)
LASSO	50	0.926 (0.041)	0.913 (0.062)	0.940 (0.053)
LDA	72	0.926 (0.039)	0.921 (0.065)	0.932 (0.055)
QDA	46	0.930 (0.025)	**0.931** (0.038)	0.928 (0.038)
KNN	76	0.774 (0.041)	0.723 (0.067)	0.827 (0.057)
First Derivative				
Method	FPCs	Accuracy	Sensitivity	Specificity
LR	46	0.921 (0.031)	0.916 (0.047)	0.926 (0.062)
RIDGE	36	0.922 (0.027)	0.908 (0.056)	0.936 (0.046)
ENET	34	0.925 (0.024)	0.918 (0.046)	0.933 (0.041)
LASSO	34	0.924 (0.025)	0.918 (0.046)	0.93 (0.045)
LDA	34	0.919 (0.033)	0.908 (0.049)	0.931 (0.051)
QDA	34	**0.930** (0.031)	0.935 (0.054)	**0.926** (0.043)
KNN	30	0.928 (0.028)	**0.936** (0.029)	0.919 (0.048)
Second Derivative				
Method	FPCs	Accuracy	Sensitivity	Specificity
LR	42	0.850 (0.048)	0.851 (0.065)	0.849 (0.071)
RIDGE	42	0.854 (0.048)	0.858 (0.065)	0.851 (0.069)
ENET	292	0.870 (0.053)	0.896 (0.054)	0.844 (0.088)
LASSO	290	0.866 (0.057)	0.891 (0.058)	0.840 (0.100)
LDA	146	0.855 (0.040)	0.859 (0.058)	0.851 (0.072)
QDA	164	**0.903** (0.038)	0.858 (0.071)	**0.950** (0.034)
KNN	158	0.891 (0.040)	**0.926** (0.050)	0.854 (0.069)

Table 15. FPCA-based classifier performance summarized by accuracy, sensitivity, and specificity for all classifiers. Results are given for FPCs derived from SLE plasma thermogram original curves and their first and second derivatives. The number of FPCs that returns the highest mean test set accuracy is given along with the mean and standard deviation for each metric.

accuracy of 92.1% when using 46 FPCs for the first derivative curves, and 85.0% when using 42 FPCs for the second derivative curves.

Learning algorithm boxplots show a different behavior for the LASSO classifier. Classification performance based on LASSO improves upon increasing the number of FPCs included up to 50 components. However, because of the selection properties of LASSO, there is minimal loss in performance as the number of FPCs continues to increase. Hence, LASSO is capable of producing high accuracy classifiers even when large numbers of FPC scores are used as predictors. Similar behaviors are found for the other penalized-LR methods, RIDGE and ENET, which are summarized in Appendix A: Figure A1.

The penalized-LR methods RIDGE, ENET, and LASSO produce mean test set accuracies of 93.5%, 93.2%, and 92.6% respectively, when based on FPC scores calculated from the SLE FDO original curves. Each of these classifiers obtains maximum accuracy upon the inclusion of the first 50 FPCs. The same classifiers return 92.2%, 92.5%, and 92.4% when using the first 34 FPCs of the first derivative curves. When using FPCs from second derivative curves slight drops in performance were found, with mean test set accuracies of 85.4%, 87.0%, and 86.6%, respectively. There is also less agreement when using the second derivative FPCs, with RIDGE accuracy maximized when using 42 FPCs, while ENET and LASSO used the first 292 and 290 FPCs.

These classification performances are significantly improved from the analysis presented in Chapter 4, which provided mean test set accuracies only as high as 91.6% when using RIDGE on first derivative predictors. Conventional ML estimates as well as

penalized-LR both show significant improvements when using FPC score predictors. The learning algorithms were also incorporated using LDA, QDA, and KNN. Learning algorithms using LDA (not shown) had a similar behavior and performance to that of the penalized-LR classifiers. LDA achieves mean test set accuracies of 92.6%, 91.9%, and 85.5% for original, first derivative, and second derivative FPCs, respectively. These accuracies are maximized when using the first 72, 34, and 146 FPCs.

Summarizing plots for the learning algorithms using QDA and KNN are given in Appendix A: Figure A2. QDA shows the most complex behavior when iterating over the number of included FPCs. QDA shows the same sharp increase in performance near 50 FPCs, but also has drastic performance loss when more than 250 FPCs are used in classifier estimation. This is because the variance-covariance estimation becomes unstable when the number of FPCs included is too large. Although this is limiting, QDA could be implemented into the learning algorithms unlike the contemporary investigation of Chapter 4. QDA performance improves in general from that of LDA, resulting in mean test set accuracies of 93.0%, 93.0%, and 90.3% for FPC scores derived from original curves and its first and second derivatives, respectively. These accuracies are maximized when using the first 46, 34, and 164 FPCs.

The final classifier, KNN, is the most insensitive to increasing the number of FPCs. Classification accuracy is low when only a few FPCs are used, but reaches optimal levels of performance at a much smaller number of included FPCs. Additionally, KNN is nearly unaffected as the number of FPCs grows, returning equivalent mean test set accuracies for nearly all values of k. KNN classification performance is reduced from

that of the other classifiers when using FPCs based on original curves: KNN only achieves a mean test set accuracy of 77.4% when using the first 76 FPCs. This improves when FPCs from first derivative curves are considered instead: a mean test set accuracy of 92.8% is returned when using the first 30 FPCs. Finally, KNN achieves a mean test set accuracy of 89.1% when using the first 158 FPCs derived from second derivative curves. This is a similar result to that found in Chapter 4, with KNN performing better on derivative predictors.

FPCA-based classifiers can also be evaluated using the naïve ensemble and weighted ensemble strategies developed in Chapter 5. Ensembles of the top classifiers from each derivative order based on the FPC algorithms presented above were investigated. The results of naïve ensembles of the FPCA-based classifiers from all three curves are summarized in Table 16. Results are shown for all seven classifiers, with a general trend that naïve ensembles improve classification performance over using a single set of curves. LR is unchanged from a mean test set accuracy of 92.1% found from FPCA analysis of first derivative curves. RIDGE and LDA suffer minor loses in comparison with using any single derivative order, returning naïve ensemble accuracies of 92.4% and 91.9%, respectively.

The naïve ensembles from ENET, LASSO, QDA, and KNN all improve over the use of any single derivative order. Each of the four methods achieves mean test set accuracies of 93.0% or higher, with QDA achieving the highest accuracy rate of 93.8%. The resulting ensembles have unique behavior in comparison with the use of standard predictors in Chapter 4. FPCA-based classifiers produce high specificity models, with

Naïve Ensemble			
Method	Accuracy	Sensitivity	Specificity
LR	0.921 (0.030)	0.920 (0.032)	0.923 (0.059)
RIDGE	0.924 (0.037)	0.914 (0.047)	0.935 (0.053)
ENET	0.936 (0.027)	**0.935** (0.037)	0.938 (0.048)
LASSO	0.930 (0.033)	0.929 (0.044)	0.931 (0.053)
LDA	0.919 (0.037)	0.911 (0.053)	0.926 (0.058)
QDA	**0.938** (0.020)	0.928 (0.045)	**0.949** (0.026)
KNN	0.933 (0.029)	0.928 (0.044)	0.938 (0.043)

Table 16. Naïve ensemble classification results using optimized FPCA-based classifiers. Results are shown for all seven classifiers investigated, with performance summarized by accuracy, sensitivity, and specificity for all classifiers. Summary metrics are given as mean with standard deviations in parentheses.

mean specificity as high as 94.9% for QDA naïve ensembles. Standard predictors produced naïve ensembles with mean accuracies only as high as 91.9%, which were primarily driven by high sensitivity to SLE (94.3%). This is an interesting switch in behavior when classifiers are instead built based on FPCA.

Weighted ensemble strategies were also considered using the optimized FPCA-based classifiers. The results of using equally-weighted and accuracy-weighted ensembles are summarized in Table 17. Equally-weighted ensembles show small changes in classification performance when using mixtures of original curves with first derivative classifiers or both first and second derivative classifiers. QDA produces the highest equally-weighted ensemble mean accuracy of 93.6% when using only original curves with their first derivatives. Equally-weighted ensembles show an interesting drop in classification performance when using original curves combined with second derivatives, or mixtures of first and second derivative classifiers. When classifiers from all three derivative orders are equally weighted, QDA can achieve 93.8% mean test set accuracy.

Interesting differences are observed when ensemble mixtures are weighted by the resulting accuracy of the separate classifiers. QDA produces an equivalent performance of 93.6% when using the combination of original curve classifiers with first derivative classifiers. Strikingly, QDA mean test set performance spikes to 94.0% when using an accuracy weighted ensemble of original classifiers and second derivative classifiers. This is an improvement of 3.8% when using accuracy weighted instead of equally weighted ensemble strategies. This result demonstrates that ensemble models are sensitive to the

Equally Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.924 (0.025)	0.904 (0.028)	0.903 (0.040)	0.923 (0.036)
RIDGE	0.926 (0.039)	0.901 (0.038)	0.902 (0.029)	0.911 (0.035)
ENET	0.930 (0.033)	0.919 (0.039)	0.913 (0.034)	0.924 (0.038)
LASSO	0.932 (0.034)	**0.924** (0.033)	0.914 (0.033)	0.920 (0.035)
LDA	0.928 (0.043)	0.896 (0.040)	0.893 (0.039)	0.912 (0.034)
QDA	**0.936** (0.024)	0.912 (0.042)	**0.915** (0.044)	**0.938** (0.020)
KNN	0.889 (0.028)	0.923 (0.032)	**0.915** (0.033)	0.937 (0.035)
Accuracy Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.926 (0.028)	0.906 (0.028)	0.905 (0.041)	0.923 (0.036)
RIDGE	0.926 (0.039)	0.908 (0.038)	0.904 (0.029)	0.913 (0.036)
ENET	0.930 (0.033)	0.919 (0.041)	0.913 (0.033)	0.925 (0.040)
LASSO	0.931 (0.033)	0.919 (0.037)	0.915 (0.034)	0.923 (0.040)
LDA	0.928 (0.042)	0.903 (0.044)	0.904 (0.034)	0.914 (0.035)
QDA	**0.936** (0.026)	**0.940** (0.030)	**0.934** (0.030)	**0.939** (0.021)
KNN	0.897 (0.025)	0.919 (0.030)	0.916 (0.025)	0.938 (0.031)

Table 17. Weighted ensemble classification results using optimized FPCA-based classifiers. The ensemble probabilities for the combination of all derivative orders (D^0 : original curve, D^1 : first derivative, and D^2 : second derivative) are given. Performance is summarized by accuracy with the test set mean and standard deviation recorded in parentheses.

weighting coefficients. The resulting QDA mean test set accuracies rival that of the final ESFuNC models presented in Chapter 6. The top performing ESFuNC ensemble was produced from the combined ensemble strategy using normal-kernel PW classifiers returned a mean test set accuracy of 94.3%.

Investigation of FPCA-based supervised learning algorithms provides a solution to overcoming many of the difficulties encountered when using standard predictors in Chapter 4. Using FPC scores as predictors alleviates multicollinearity within the data improving the classification performance of nearly all classifiers investigated within this dissertation. Ensemble strategies that combine FPCA-based classifiers from original curves with its first and second derivatives also demonstrated improved classification performance.

Importantly, this section also highlighted the necessity of validating the number of FPCs included as predictors in LR models. Although FPCs are typically included based on explanation of a high percentage of the variability within the data, this should not be a limiting criterion during model building. By using only the first six components from PCA analysis, (Garbett and Brock 2016) suggested that PCA-based LR returned sub-par accuracy rates of only 71% for SLE classification. Although using a small number of FPCs does result in low classification accuracy, FPCA-based learning algorithms are able to suggest an optimized number of FPCs to include as predictors. Classification performances improve dramatically upon validation the number of FPCs used, producing classifiers that are capable of discriminating between SLE and non-SLE alternatives with accuracy rates as high as 94.0%.

7.3 Ensemble of Per-Pixel Classifiers

Several ensemble strategies have been considered within this dissertation. The results presented in Chapters 5 and the FPCA-based classifiers studied above demonstrated that simple ensembles using classifiers based on multiple derivative orders can improve classification performance. The ESFuNC algorithm developed in Chapter 6 introduced more complex ensemble strategies based on using weighted combinations of estimated class probabilities. FSS and BSS methods were designed to choose segmented-FDOs that optimized the accuracy of the final ensemble. One difficulty common to these results has been how to choose weighting constants incorporated into the ensemble strategies.

This section introduces designs for estimating ensemble combinations using LR methods. All R code developed for PPCs is maintained in the Github repository (https://github.com/BuscagliaR). As a limiting behavior to segmented-FDO ensembles used in the ESFuNC algorithm, the idea of PPCs is also introduced. A PPC refers to the estimation of unique classifiers at each separate predictor. For SLE plasma thermograms, this corresponds to producing classifiers at each of the 451 unique temperatures represented in the SLE FULL predictor set. Each of the classifiers studied within this dissertation produces estimated class probabilities. Classifiers can thus be produced at each unique predictor providing a set of estimated class probabilities that can be used for ensemble model construction.

SLE plasma thermogram PPCs were estimated using contemporary KNN methods as introduced previously. This choice was made based on the success of NCs observed in

earlier chapters, but per-pixel classification can be performed with any of the classifiers introduced thus far. The corresponding DSC signal output at each temperature within the SLE FULL predictor set was used to determine the Euclidean distance between all 589 samples. This produces a set of 451 PPCs and their resulting estimated class probabilities. A learning algorithm was implemented to validate the nonparametric tuning constant for each pixel by maximizing the LOOCV accuracy.

The LOOCV accuracies obtained at each SLE plasma thermogram pixel for the original curves along with first and second derivatives are shown in Figure 16. There is remarkable structure to the per-pixel results when using the original curves. A maximum LOOCV accuracy of 74.2% is achieved at a temperature of 65.6 ^{o}C, which corresponds to the temperature where the mean SLE and non-SLE curves have the largest difference (Figure 1). This is not surprising as the KNN classifiers are based on distance metrics, and the mean curves reflect where the major regions of difference should be located. First derivative PPCs also retain minimal structure, with peaks occurring near 62 and 68 ^{o}C. Although noisy, the first derivative PPCs obtain a maximum LOOCV accuracy of 74.4%. Second derivative classifiers show no definitive structure in the per-pixel accuracies with generally low performance and a maximum accuracy of only 59.8%.

The goal of per-pixel analysis is to use the estimated class probabilities to produce high accuracy ensembles from combinations of PPCs; although no PPC achieves outstanding classification performance, PPEs may be capable of boosting performance. There are several published strategies for the combination of classifiers (Freund and Schapire 1995; Ho et al. 1994). Several strategies have already been introduced,

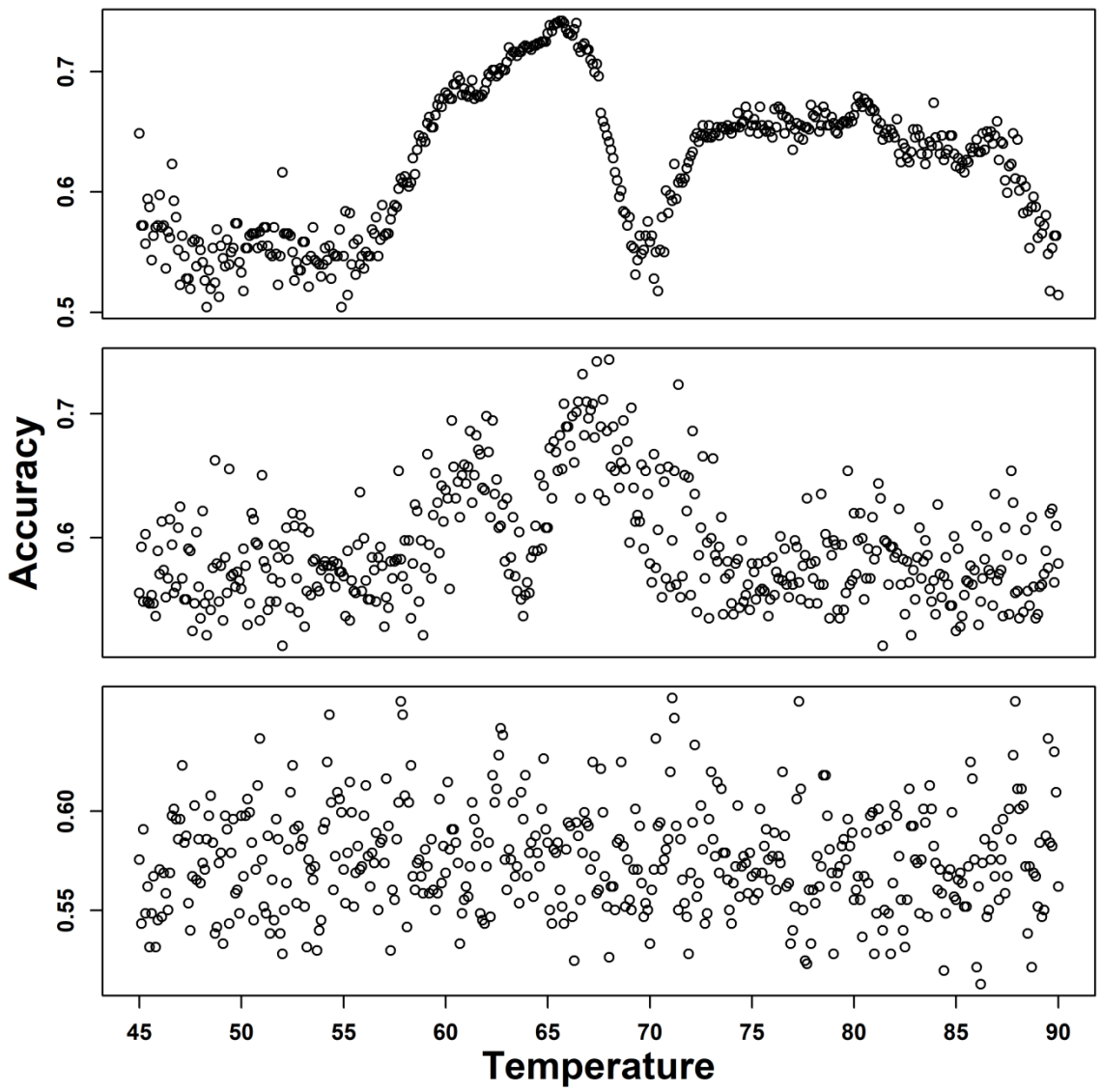

Figure 16. LOOCV accuracies resulting from PPCs for the SLE plasma thermogram data set. Accuracies were obtained through the validation of KNN classifiers at each temperature. Accuracies are shown for classifiers based on original curves (top), first derivatives (middle), and second derivatives (bottom).

including naïve voting schemes and simple weighted combinations of estimated class probabilities. These methods were found in Chapter 5 to provide significant boosts to classification performance when classifiers based on multiple derivative orders were mixed.

As a starting point, PPCs were combined using naïve class voting and weighted ensemble strategies (Section 5.6). Naïve PPEs were constructed separately for each set of PPCs collected from original, first, and second derivatives producing LOOCV accuracies of 75.4%, 86.6%, and 72.7%. Accuracy increases over using any single pixel by more than 10% for first and second derivative based naïve PPEs, demonstrating the impact ensembles can have on boosting performance. Naïve PPEs were also produced using combinations of PPCs from all three curves. When doing so, a LOOCV accuracy of 88.5% can be achieved when using estimated classes from first and second derivative PPCs.

Weighted PPEs use estimated class probabilities rather than estimated classes. Equally-weighted PPEs return LOOCV accuracies of 74.4%, 80.5%, and 75.7% when based on original, first derivative, and second derivative PPCs. A maximum LOOCV of 85.7% can be obtained when using an equally-weighted ensemble of first and second derivative PPCs, similar to naïve voting. Slight reduction in PPE performance is observed when using accuracy-weighted ensembles, with the best accuracy-weighted PPE producing an 85.2% accuracy rate for a mixture of first and second derivative PPCs.

PPCs using the simple ensemble techniques are capable of achieving LOOCV accuracy as high as 88.5%. This demonstrates that using classifier estimates can produce

effective predictions, even when individual PPCs are under-performing. The methods studied above suffer from two significant factors. First, no selection of PPCs is done prior to producing PPEs. It may be that pruning or removing of certain PPCs could improve results. Second, when weighting constants are used, they are only estimated from PPC accuracies, and no fitting is done to attempt to optimize weighting constants.

To determine if selection could improve PPC performance, the FSS routines developed for the ESFuNC algorithm were employed. Although FSS was constructed to accept segmented-FDO classifier information, the same algorithms can be used to evaluate stepwise ensembles of PPCs. Separate stepwise PPEs were constructed using PPCs from each derivative order. In addition, all combinations of derivative orders were also considered. This was possible due to the computational speed of the FSS algorithm, which efficiently returns stepwise PPEs even when using all 1353 PPCs.

The FSS algorithm was used allowing for both equally-weighted and accuracy-weighted ensembles. A maximum LOOCV accuracy of 85.9% was achieved from the equally-weighted method, which used only first derivative PPCs. The final ensemble selects 8 of the possible 451 classifiers. The accuracy-weighted PPE was able to achieve 87.1% LOOCV accuracy, and did so through the combination of first and second derivative PPCs. The final ensemble uses only 12 of the potential 902 PPCs. Although FSS accuracies are slightly lower than that of the naïve PPEs, the results show the importance of using selection methods. Nearly equivalent results can be obtaining when using only a small number of the PPCs.

These results lead to the conception of producing PPEs based on LR. Let the estimated class probabilities predicted from the kth PPC be denoted P_k. The PPEs considered so far evaluate a linear combination of the estimated class probabilities, producing PPE estimated class probability P_E. This can be written as

$$P_E = \sum_k \beta_k P_k$$

where in Chapter 5 restrictions were introduced on the weighting factors such that

$$\sum_k \beta_k = 1.$$

If we consider the PPC estimated class probabilities as predictors in a LR model, we can rewrite the estimation of the PPE estimated class probabilities as

$$\log\left(\frac{P_E}{1-P_E}\right) = \beta_0 + \sum_k \beta_k P_k.$$

This suggests that PPEs could be produced by using LR estimates for the weighting coefficients. Such methods have been used previously with successful improvements to classification performance, but considered mixing of classifiers from alternative methodologies (Ho et al. 1994). Introduced in this work is the idea of PPEs based on LR-estimated weights.

LR estimated PPEs are evaluated rapidly as a natural extension of the learning algorithms developed in Chapter 4. PPC estimated class probabilities were incorporated into the learning algorithms as predictors. PPEs were produced using ML estimates (LR), along with the penalized-LR methods, LASSO, ENET, and RIDGE. As an

additional means of cleaning the PPC predictor sets, variance VIF were considered (Craney and Surles 2002). VIF is defined as

$$VIF = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination for a linear regression model fitting the ith predictor using all remaining predictors.

If a predictor can be explained by a linear combination of the remaining predictors, R_i^2 will approach 1, causing VIF to increase. Pruning is conducted by removal of the predictor having the largest VIF. An iterative VIF algorithm was used to prune the PPC predictor set iteratively, removing a single PPC at a time until the remaining predictors each have VIFs that fall below a predetermined threshold. For this study, VIF thresholds of 20, 10, and 5 were used, corresponding to R^2 of 0.95, 0.9, and 0.8, respectively. This was done in an attempt to see if a particular VIF threshold could produce improvements in the LR estimated PPEs.

VIF pruning of the PPC estimated class probabilities resulted in different pruning based on which derivative order was considered. VIF pruning of original curve PPC estimated class probabilities resulting in removal of 182, 225, 262 predictors using thresholds of 20, 10, and 5, respectively. The first and second derivative predictor sets were less pruned, with removal of only 91 and 69 predictors at a VIF threshold of 5. This indicates significantly less collinearity between derivative-based estimated class predictors in comparison with predictors constructed from original curve pixels.

The classification performances of PPEs estimated using LR are summarized in Table 18 for the unpruned predictor sets as well as the resulting predictors after VIF pruning. The resulting LR-estimated PPEs have remarkable resemblance to the contemporary analysis conducted in Chapter 4. VIF is found to produce no significant improvements to the overall classification performance of LR-estimated ensembles. PPEs based on RIDGE are the highest performing, with a mean test set accuracy of 91.4% being obtained when first derivative PPCs are considered. This is nearly equivalent to the 91.6% accuracy rate achieved during the contemporary analysis when using RIDGE (Table 1).

Combining the estimated class probabilities from original, first derivative, and second derivative PPCs was also considered. Because VIF pruning did not improve classification performance, all PPCs were used when producing combined predictor sets. The results of RIDGE, ENET, and LASSO estimated PPEs are presented in Appendix A: Table A9. Combined per-pixel predictor sets produce minimal improvements to classification performance. PPEs estimated from RIDGE show a loss in performance, while ENET is capable of achieving an accuracy rate of 90.5% when using a combination of estimated probabilities from first and second derivative PPCs. This results matches the contemporary investigation, where combining predictors from multiple derivative orders had minimal influence on classification performance.

Improved from using combined predictor sets is the use of the ensemble strategies discussed in Chapter 5. Naïve and weighted ensembles were considered for mixing of LR-estimated PPEs from each derivative order. The results of the naïve ensemble of

Original Curves				
Method	Unpruned	$VIF = 20$	$VIF = 10$	$VIF = 5$
LR	0.632 (0.075)	0.672 (0.057)	0.692 (0.051)	0.699 (0.050)
RIDGE	**0.811** (0.045)	**0.809** (0.055)	**0.814** (0.057)	**0.805** (0.048)
ENET	0.806 (0.053)	0.802 (0.053)	0.811 (0.050)	0.799 (0.056)
LASSO	0.802 (0.051)	0.800 (0.048)	0.805 (0.046)	0.796 (0.053)
First Derivative				
Method	Unpruned	$VIF = 20$	$VIF = 10$	$VIF = 5$
LR	0.710 (0.065)	0.672 (0.057)	0.692 (0.051)	0.699 (0.050)
RIDGE	**0.914** (0.032)	**0.805** (0.059)	**0.817** (0.057)	**0.805** (0.049)
ENET	0.894 (0.035)	0.802 (0.055)	0.809 (0.047)	0.800 (0.053)
LASSO	0.882 (0.047)	0.800 (0.053)	0.803 (0.046)	0.794 (0.051)
Second Derivative				
Method	Unpruned	$VIF = 20$	$VIF = 10$	$VIF = 5$
LR	0.694 (0.075)	0.672 (0.057)	0.692 (0.051)	0.699 (0.050)
RIDGE	**0.871** (0.052)	**0.805** (0.057)	**0.817** (0.052)	0.805 (0.048)
ENET	0.858 (0.063)	0.802 (0.051)	0.809 (0.046)	**0.806** (0.050)
LASSO	0.846 (0.057)	0.803 (0.050)	0.803 (0.047)	0.795 (0.054)

Table 18. LR-estimated PPE classification performances. 10-fold CV was performed using PPC estimated class probabilities as predictors. Results are shown for the full predictor sets along with pruned predictor sets using VIF thresholds of 20, 10, and 5. Classification performance is reported by test set mean and standard deviation.

PPEs are presented in Table 19. Naïve ensembles of PPEs show general improvements from using any single LR-estimated PPE. Classification performance is improved when PPEs constructed from original, first, and second derivative PPCs are mixed. A naïve ensemble of RIDGE-estimated PPEs from each separate curve obtains a mean test set accuracy of 92.8%.

Naïve ensemble of LR-estimated PPEs produce classification performance improved in comparison with contemporary results (Table 10). Naïve ensembles using standard predictor methods produced models with accuracy rates only as high as 91.9% when using KNN classifiers, and only 91.4% when using penalized-LR classifiers. This result shows that PPCs have potential for improving classification performance over standard predictors. Unique to note is that naïve ensembles using contemporary predictor methods were primarily driven by high sensitivity to SLE, while the naïve ensembles resulting from PPE combinations are driven by high specificity. If conventional predictor ensembles and PPEs can be mixed to further improve classification performance has yet to be studied.

The weighted ensemble strategies were also employed using the resulting LR-estimated PPEs. Mixtures involving all combinations of original, first, and second derivative PPEs were considered. The results of the equally-weighted and accuracy-weighted ensembles are presented in Table 20. Remarkably, mixtures of PPEs from multiple derivative orders are capable of producing classifiers with accuracy rates as high as 93.6%, resulting from RIDGE-estimated PPEs. Equally-weighted ensembles of LR-

Naïve Ensemble			
Method	Accuracy	Sensitivity	Specificity
LR	0.751 (0.062)	0.765 (0.070)	0.735 (0.087)
RIDGE	**0.928** (0.025)	0.904 (0.051)	**0.952** (0.037)
ENET	0.909 (0.042)	**0.906** (0.069)	0.913 (0.050)
LASSO	0.905 (0.044)	0.898 (0.067)	0.913 (0.056)

Table 19. Naïve ensemble classification performances using LR-estimated PPEs. PPEs were generated separately using original, first, and second derivative PPCs. Ensembles were then constructed by mixing of all three PPEs. Classification performance is reported by test set mean and standard deviation.

Equally Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.703 (0.063)	0.689 (0.062)	0.745 (0.070)	0.750 (0.060)
RIDGE	**0.908** (0.034)	**0.914** (0.040)	**0.929** (0.033)	**0.934** (0.025)
ENET	0.896 (0.041)	0.900 (0.046)	0.915 (0.038)	0.924 (0.042)
LASSO	0.884 (0.042)	0.902 (0.041)	0.903 (0.051)	0.917 (0.045)
Accuracy Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.721 (0.060)	0.718 (0.055)	0.757 (0.051)	0.749 (0.060)
RIDGE	**0.911** (0.032)	**0.914** (0.041)	**0.930** (0.031)	**0.936** (0.023)
ENET	0.896 (0.039)	0.902 (0.040)	0.916 (0.037)	0.925 (0.042)
LASSO	0.886 (0.045)	0.902 (0.043)	0.904 (0.049)	0.917 (0.044)

Table 20. Equally- and accuracy-weighted ensemble classification performances using LR-estimated PPEs. All combinations of original (D^0), first derivative (D^1), and second derivative (D^2) PPEs were considered. Classification performance is reported by mean and standard deviation test set accuracy.

estimated PPEs based on original, first, and second derivatives achieve 93.4% accuracy rates when using RIDGE estimation of the separate PPEs.

PPCs provide a novel approach to producing ensembles models based on classifier information from separate predictors. Unlike conventional predictors, PPCs were used to produce a new set of predictors based on LOOCV estimated class probabilities. Investigation of the ensemble strategies used throughout this dissertation demonstrated that naïve voting and simple linear combinations of PPCs could only produce classifiers with accuracy rates as high as 88.5%. LR was incorporated to estimate ensembles leading to improved PPE performance; mixing of PPEs from multiple derivative orders boosts classification accuracy rates to as high as 93.6%. These accuracy rates are in the neighborhood of the classifiers found from the ESFuNC algorithm and FPCA-based classifiers. Importantly, using PPC estimated class probabilities as predictors improves performance in comparison with the conventional predictors presented in Chapter 4. This suggests that building classifier sets, such as PPCs, could have unique implications on classification performance.

7.4 Fused-LASSO Estimation of Per-Pixel Ensembles

The final section investigates penalized-LR estimation of the PPEs using fused-LASSO. Fused-LASSO is a penalization method similar to LASSO and RIDGE discussed in Chapter 4. The penalized likelihood problem for fused-LASSO LR can be formulated as:

$$\frac{1}{N}\sum_{i=1}^{\mathrm{N}}\left[\mathrm{y_i}\left(\beta_0 - \mathrm{x_i^T}\beta\right) - \log\left(1 + \mathrm{e}^{\left(\beta_0 - \mathrm{x_i^T}\beta\right)}\right)\right] + \lambda_1 \sum_{\mathrm{j}=1}^{\mathrm{p}}\left|\beta_j\right| + \lambda_2 \sum_{\mathrm{j}=2}^{\mathrm{p}}\left|\beta_j - \beta_{j-1}\right|.$$

Unique to the fused-LASSO classifier is the l_1-penalization of neighboring coefficients in addition to l_1-penalization of the coefficients. Solutions to the fused-LASSO LR problem are difficult to compute due to the non-smooth and non-separable penalties (Liu et al. 2010). Several studies have investigated solutions to the fused-LASSO problem with focus having been primarily on a least squares loss functions (Tibshirani et al. 2005); the logit loss function, used for LR, is less studied in the literature. Nevertheless, there does exist several unique fused-LASSO solvers implemented using different gradient descent (Liu et al. 2010), iterative searches (Lee et al. 2014) and Newton-Raphson based approaches (Goeman et al. 2012).

Difficulties in the optimization algorithms for fused-LASSO are mainly related to the validation of λ_1 and λ_2 tuning parameters. The R package **penalized** provides efficient solutions for the fused-LASSO classifiers using both gradient ascent and Newton-Raphson approaches (Goeman et al. 2012; Goeman 2010). For the estimation of PPEs, solutions to the fused-LASSO classifier at a given λ_1 and λ_2 are typically found efficiently and with minimal computational time. However, it was found that certain combinations of λ_1 and λ_2 could lead to significant jumps in computational times for SLE plasma thermogram classification. Computational times as long as 16 hours were recorded for the evaluation of a single KCV fold. Although the package includes KCV routines for finding optimized λ_1 and λ_2 tuning constants, it was determined that evaluation of a predefined grid resulted in computationally feasible solutions.

The goal of using fused-LASSO was to determine if regions of pixels could be grouped to gain information about which regions of the temperature domain are critical for discriminating SLE vs. non-SLE alternatives. The fused-LASSO provides two unique properties when changing λ_1 and λ_2 tuning constants. The constant λ_1 controls the sparsity of the predictors included in the final classifier: as λ_1 increases, more predictors are driven to have coefficients of zero, essentially excluding them from the classifier. The λ_2 constant affects the variability observed by neighboring coefficients. Large values of λ_2 enforce neighboring coefficients to have nearly equal magnitudes, effectively grouping important predictors into unique regions. Fused-LASSO coefficient grouping may suggest not only which predictors are important for classifier performance, but also what regions of the domain are critical for population discrimination. For data sets where predictors have inherent structures, such as the temperature grid on which SLE plasma thermograms are collected, providing regions of importance can have practical implications for practitioners.

To alleviate the computational issues encountered from λ_1 and λ_2 optimization, learning algorithms were developed to produce fused-LASSO classifiers under the restriction of $\lambda_1 = \lambda_2$. The classification performances of fused-LASSO estimated PPEs using original, first, and second derivative PPCs are summarized in Figure 17. The figure gives boxplots of classification accuracy for each λ considered in the learning algorithm. A λ grid of 0.01 to 100 was used with 10 values sampled per log. The figure shows that a null model is returned when λ is large, providing accuracy rates near 50%. As the tuning constants are decreased, nearly equivalent classification performances are found over

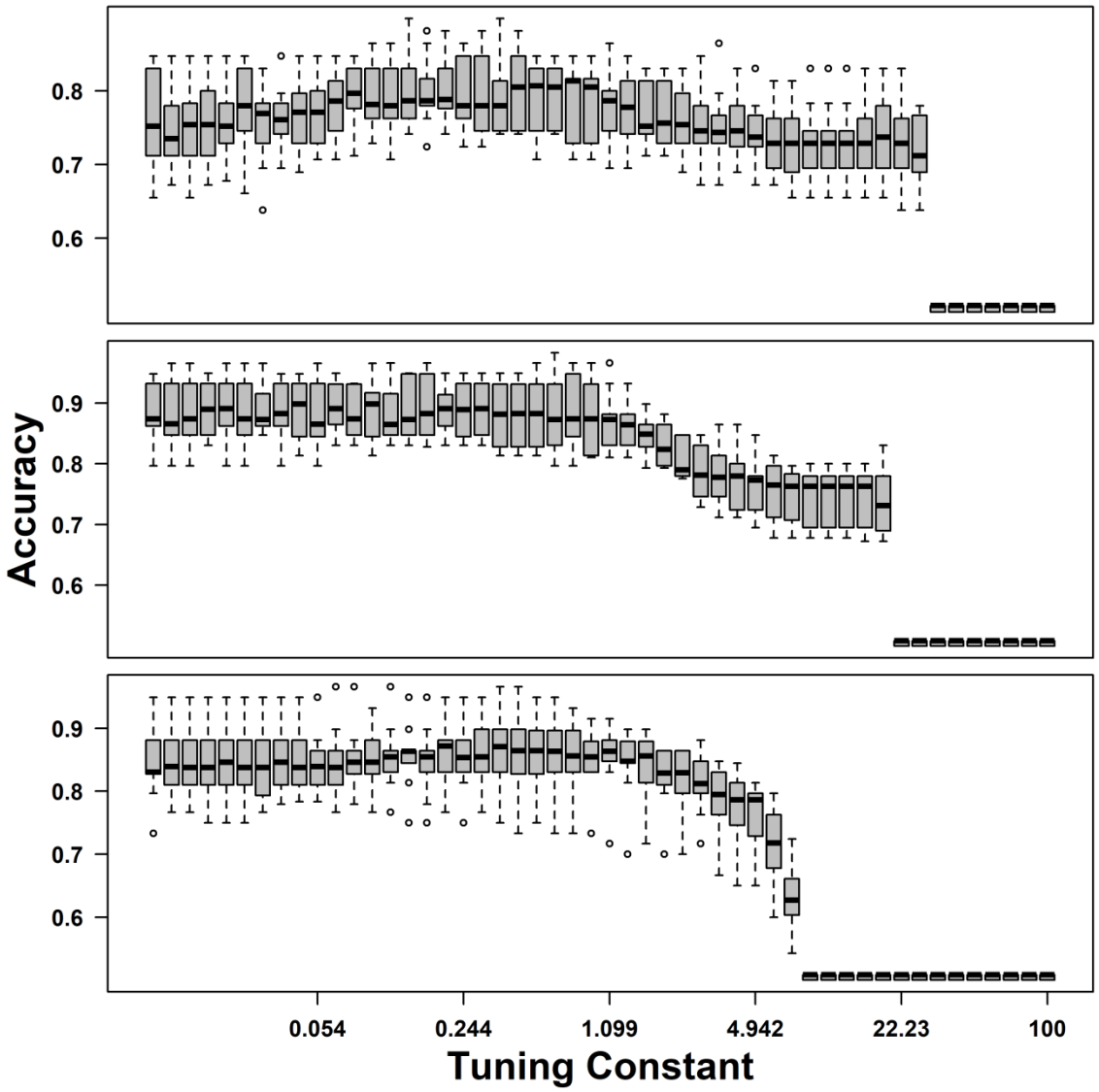

Figure 17. Tuning constant validation for fused-LASSO estimation of PPEs under the constraint of $\lambda_1 = \lambda_2$. CV is summarized by accuracy boxplots given at each λ tested. Results are shown for PPEs based on original curve (top), first derivative curve (middle), and second derivative curve (bottom) PPCs.

several orders of magnitude for λ. Maximum mean test set accuracies of 80.5%, 89.3%, and 86.4% are returned from fused-LASSO estimated ensembles based on original, first, and second derivative PPCs, respectively.

The naïve and weighted ensemble strategies were applied to the fused-LASSO estimated PPEs. Naïve ensembles based on PPE estimated classes from all three curves produced a mean test set accuracy of 90.7%. This result corresponds well to other naïve ensembles using LASSO classifiers; conventional result and PPE returned accuracy rates of 90.6% (Table 10) and 90.5% (Table 19). All combinations of fused-LASSO estimated PPEs were also evaluated using equally-weighted and accuracy-weighted ensembles; results are equivalent to that of previous LASSO based classifiers. An equally-weighted mixture of all three PPE classifiers returned a 91.5% accuracy rate, while the accuracy-weighted ensemble returned 91.3%. These results agree well with conventional methods (Table 10) and PPEs (Table 19). The fused-LASSO classifier is thus capable of returning nearly equivalent classification results to that of using LASSO.

Equally important as the estimation of high accuracy PPEs is the evaluation of pixel grouping based on the fused penalty terms. Figure 18 provides an illustration of how fused-LASSO coefficients change with the magnitude of λ. The figure gives the resulting model coefficients from fused-LASSO LR using PPCs based on original curves. Coefficients are dispersed and resemble that of estimated coefficients from standard ML estimation of LR classifiers when λ is small. The figure demonstrates that as λ increases coefficients shrink to zero, allowing for selection of important pixels for classification. This is an equivalent behavior to the LASSO classifier studied earlier, but now, groups of

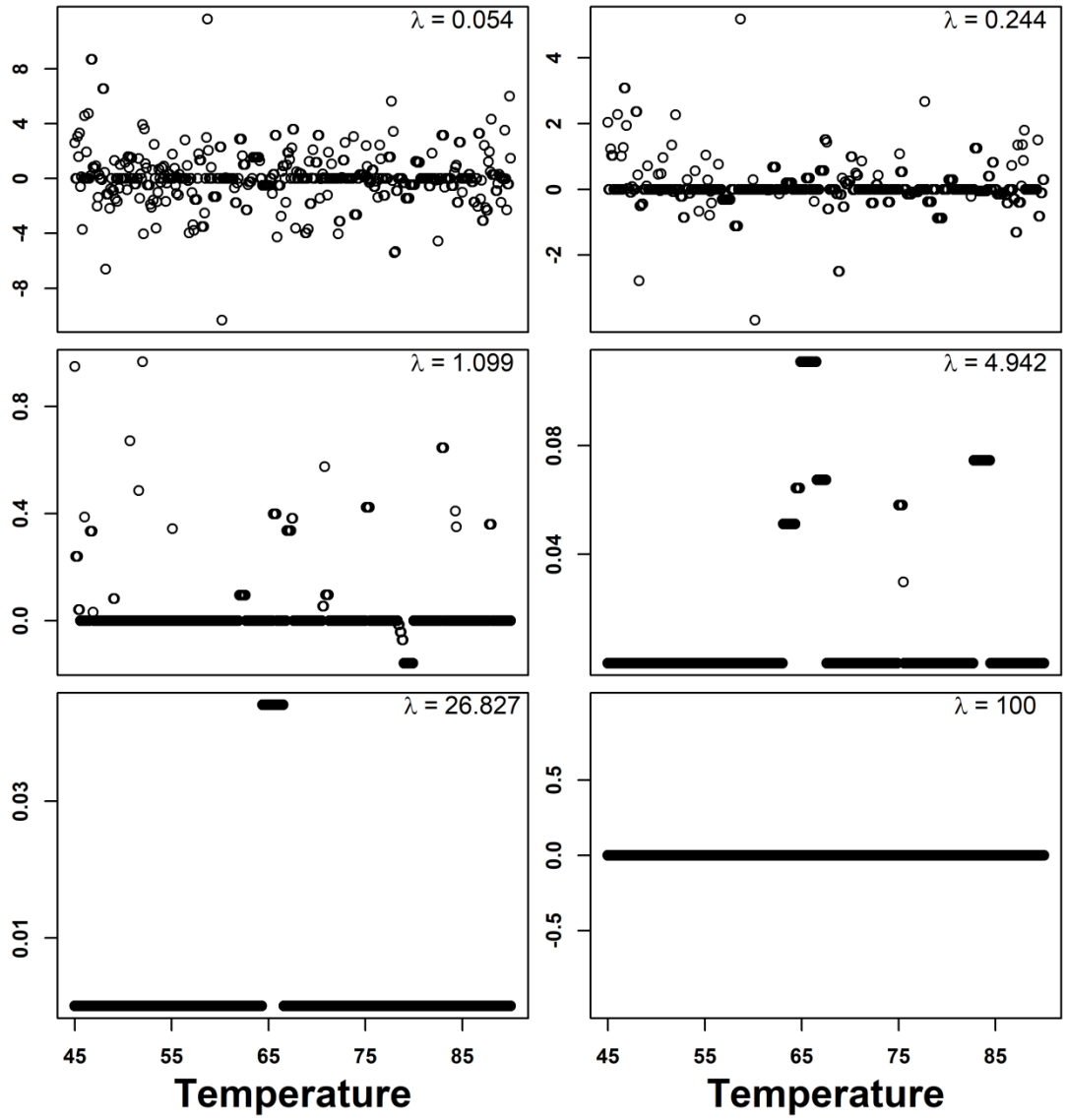

Figure 18. Coefficients estimated by fused-LASSO for PPEs based on original SLE plasma thermogram curves. Shown are six different λ values given in the top right corner of each panel. Vertical axis gives the coefficient magnitudes. All fused-LASSO results were produced using $\lambda_1 = \lambda_2$.

coefficients are set to have equal magnitudes. When $\lambda = 4.942$, distinct regions of equal coefficient magnitudes are clearly visible; these regions correspond to temperatures which are important for discriminating SLE vs. non-SLE alternatives. Further studies may investigate if such regions can be useful in understanding the biochemical processes behind the discrimination of SLE vs. non-SLE alternatives.

7.5 Conclusions

The analysis of SLE plasma thermograms was extended through analysis of FPCA-based classifiers and the development of PPEs. FPCA-based classification shows improvements over the use of original predictor values. Using FPC scores can improve classification performance, but requires the selection of how many FPC components should be used as predictors. A learning algorithm was developed that allowed for optimization of the number of FPCs included. Accuracy-weighted ensembles of QDA classifiers based on FPC scores returns mean test set accuracies of 94.0% when based on original curve and second derivative FPCs. This accuracy is nearly equivalent to that found from the ESFuNC algorithm. Alternative approaches have now been produced to achieve classification performances of 94.0% accuracy.

PPCs were then constructed as a unique take on predictor based methodologies. PPCs were used to produce estimated class probabilities from each separate predictor. For the SLE plasma thermograms, PPCs were produced at each observed temperature for original curves and their first and second derivatives using KNN classifiers.

Nonparametric tuning constants were optimized by learning algorithms for each PPC. Optimized PPCs were used to produce LOOCV estimated class probabilities that serve as predictors for construct of PPEs. This method is not limited to the use of KNN for preliminary LOOCV estimates, and other classifiers should be investigated in future studies.

The ensemble methods investigated in Chapter 5 and the FSS method developed for the ESFuNC algorithm were both tested as approaches for combining PPCs. Each of these ensemble methodologies results in relatively low-performance, achieving no higher than 88.5% using naïve ensembles of all PPCs based on first and second derivative curves. This is in part believed to be due to the choice of weighting factors. The ensemble classifiers investigated in this dissertation are based on linear combinations of estimated class probabilities. This lead to the consideration of LR based estimation of PPEs, which improved classification performance for each separate curve. Moreover, combination of PPEs based on original, first, and second derivative curves boosts performance further, achieving accuracy rates as high as 93.6% using RIDGE estimation. These classifiers approach the effectiveness of the final ESFuNC classifiers and FPCA based classifiers.

Fused-LASSO was then introduced as an estimation method capable of both predictor selection and smoothing of predictor coefficients (Tibshirani et al. 2005). Fused regression enforces the restriction that neighboring coefficients should not have large differences in magnitude. Fused-LASSO was used to create LR-estimated PPEs. A simplified search for tuning constant optimization was implemented, but resulted in

excellent insight into the potential of the fused-LASSO methodology. Classification performance of fused-LASSO estimation is equivalent with that found using LASSO, but improves interpretability of the structure of the predictor set. Fused-LASSO produces coefficient estimates that provide indications as to which regions of the argument domain are essential for classification. For SLE classification, this corresponds to temperature regions of which the estimated coefficients are smoothed to have small differences in magnitude.

These regions could be informative for numerous reasons. Biochemical based studies have been used to decompose the SLE plasma thermogram signatures into proportions of the known human plasma proteome (Garbett et al. 2009). The authors were able to reconstruct healthy patient thermograms based on combinations of thermograms collected for the individual proteins within the plasma proteome. If temperatures can be grouped during estimation of curve classifiers, this could be utilized to estimate biochemical changes in patient physiology. Such topics are of interest for future investigations, where fused-LASSO may become an important estimator for its capabilities of producing smoothed predictor coefficients.

Chapter 8

CONCLUSIONS AND FUTURE WORK

8.1 Conclusions

This dissertation has presented an in-depth statistical investigation of SLE plasma thermograms (Garbett et al. 2009). The main task was classification of patients with SLE against non-SLE alternatives using only the data provided from plasma thermograms. Topics were drawn from statistical subfields including FDA, penalized-LR, NC, supervised learning, and ensemble learning to produce novel approaches to this classification problem. FDA (Ramsay 2006) was found to be an important apparatus for a deep statistical interrogation of the thermograms as it provides flexible methods for viewing the thermograms as functions.

A conventional statistical study was conducted using discretizations of the functional representations of SLE plasmas thermograms and their first and second derivatives. Derivative curves provided information that improved population discrimination. Using penalized-LR methods, mean test set accuracies of 91.6% could be achieved using first derivative approximations (Table 1). This represents a significant improvement to recent literature reports of 89%, which required use of serological predictors in combination with thermogram data (Garbett et al. 2017). These results were achieved using fully approximated B-spline functional representations of the SLE

thermograms. Different functional approximations were tested, with the fully represented and unsmoothed SLE data providing the best classification performance.

Functional classification based on using the original SLE curves and derivative approximations as functional covariates was generally less effective than penalized-LR classifiers. Several LR estimates based on functional covariates were considered, producing accuracy rates no higher than 83.6%. NC using either discretized predictors or functional covariates were found to give promising classification results, with accuracy rates as high as 90.8% (Table 1) and 90.5% (Table 7) using only a single curve. To combine FuNC results from different functional covariates, ensemble methodologies were considered.

Ensemble learning involves mixing classifier information gained from multiple sources to improve the overall classification accuracy. Simple ensemble algorithms based on naïve voting or weighted linear combinations of estimated probabilities resulted in improvements to classification performance. Using combinations of classifiers estimated from each of the three curves considered, accuracy rates were boosted to as high as 92.8% using KNN classifiers (Table 11). This suggested NC and FuNC are effective for discrimination between SLE and non-SLE alternatives, indicating nonlinearity in the decision boundary.

This led to the development of the ESFuNC curve classification algorithm, which attempts to take advantage of the properties of FuNC. The ESFuNC algorithm estimates classifiers from multiple functional covariates that are then combined using stepwise ensemble building algorithms. FSS and BSS algorithms are introduced, with both

providing effective model building strategies with significantly different computational strategies. FSS allows for investigation of large grids and large combinations of classifiers without computational restrictions, but with limitations to the total number of combinations considered. BSS conducts exhaustive searches of all ensemble combinations and is employed when the number of classifiers considered is limited to 25 – 30.

The ESFuNC algorithm provides learning algorithms that optimize NC tuning constants and segmentation patterns for separate functional covariates. Segmented-FDOs and ensemble classifiers based on segmented-FDOs are shown through simulation to have benefits relative to conventional FuNC (Figures 10 – 12). Segmentation improves performance when populations differ only on small regions of the functional domain, while ensembles can be used to boost performance even when individual classifiers are under performing. These concepts are utilized within the ESFuNC algorithm, which also provides three unique strategies to combine FuNC resulting from multiple functional covariates.

The greedy (Figure 7), combined (Figure 8), and hierarchical (Figure 9) ensemble strategies provide unique methods for how information is mixed across functional covariates. Each of the methods has computational trade-offs, with the GES being the most computationally intensive. The GES requires that each functional covariate be optimized in tuning constant and segmentation pattern before information is combined. This in practice produces higher segmentation sizes that can lead to difficulties implementing BSS algorithms. The HES introduces an order to which the functional

covariates are evaluated. This allows each functional covariate to be optimized in the presence of information gained from earlier functional covariates.

The CES simplifies how each functional covariate is segmented. The strategy leads to simplified segmentation patterns and ensemble combinations, with low computational burden at low segmentation sizes. The computational complexity of the CES does grow in correlation with the number of functional covariates considered, with BSS only being feasible for early steps of the algorithm. Interestingly, this method is found to provide the highest mean test set accuracy for SLE classification of all classifiers studied in this dissertation: the ESFuNC algorithm achieves an ensemble classifier with 94.3% mean test set accuracy (Table 12). The classifier uses a mixture of information from original SLE plasma thermograms and their first and second derivative approximations (Table A6).

The ESFuNC algorithm was successful for improving the classification of SLE plasma thermograms. Accuracy rates of 94% suggest that thermograms have significant potential as a diagnostic technology. These results in combination with future thermogram studies hope to build a foundation for which the plasma thermogram diagnostic technique can become clinically relevant. The SLE plasma thermogram data set is also rich with additional classification problems, which will be considered in the next section.

Benchmark data sets were also investigated to show that the ESFuNC algorithm can generalize to all curve classification problems. Tecator classification results in a classifier with 99.8% mean test set accuracy (Table 13), using segments from first and

second derivatives curves. These results match well with recent publications using FDA based approaches (Li and Yu 2008), and are the first to consider an ensemble of classifier resulting from multiple derivatives. Phonetic classification was also evaluated using the ESFuNC algorithm, with classifier accuracy rates as high as 93.5% (Table 14). These results are equivalent to or higher than previous phoneme classification studies using the same phonetic speech frames. These studies combined with the SLE plasma thermograms demonstrate that the ESFuNC algorithm can generalize to a variety of curve classification problems.

FPCA-based classifiers were also considered. FPC scores can provide a predictor set free of the multicollinearity issues that interfered with the contemporary analysis of Chapter 4. A learning algorithm was developed to evaluate the number of FPCs included during classifier estimation. Weighted ensemble of classifiers based on FPCA of original, first, and second derivatives produce effective classification. Mean test set accuracies as high as 94.0% are achieved when using classifiers based on QDA, reaching performances equivalent to final ESFuNC classifiers (Table 17).

As a means of further studying the potential of ensemble learning, ensembles of PPCs were investigated. PPCs are produced by constructing unique classifiers at each separate predictor. It was chosen to base PPCs on the nonparametric KNN classifier, but any of the classifiers investigated in this work can be used. Individual pixels from PPCs achieve SLE classification accuracy rates no higher than 74.4%. PPE construction using naïve voting, weighted linear combinations of estimated class probabilities, and the FSS algorithm developed for ESFuNC were evaluated. Simplified methods achieve PPE

accuracy rates of 86.6%, which can be further boosted to 88.5% by combining the separate PPEs produced using original, first, and second derivative PPCs.

Considering the PPEs as a linear combination of PPC estimated class probabilities, it was decided to evaluate if LR estimates of PPEs could be obtained. LR estimation produces significant improvements in the final PPE accuracy, increasing to an accuracy rate of 91.4% when using RIDGE estimation based on first derivative PPCs (Table 18). If the LR based PPEs are mixed using an accuracy-weighted combination of all three curves of interest, accuracy rates are improved to 93.6% (Table 20) and begin to approach the success of the ESFuNC and FPCA-based classifiers. PPCs represent a unique methodology for constructing predictor sets, with PPEs achieving successful classification of SLE plasma thermograms. This approach shows unique potential for developing classifiers with unique boundary layer definitions, and is of significant interest for future work.

This dissertation has achieved the goal of improving SLE plasma thermogram classification. Motivated by this task, additional goals of developing novel and unique approaches to curve classification were also achieved. The ESFuNC algorithm represents a new adaptation of FuNC, which uses the combination of FDA, NC, and ensemble learning methods to produce effective classifiers. This algorithm represents a modern approach to multivariate functional data classification. The dissertation also develops fresh views to previously investigated techniques. FPCA-based learning algorithms capture high accuracy rate classifiers by removing restrictions related to how the number of FPCs is chosen. PPCs are a unique take to how predictor sets can be constructed, and

LR estimations of PPEs show promising results for producing effective classifiers. This dissertation has improved the potential clinical implications of a unique diagnostic technique while also developing novel classification approaches.

8.2 Future Work

This dissertation is filled with a litany of future directions from each of the major themes introduced. Investigations based on improving FDA classification, ESFuNC, and PPEs are immediate sources of promising work. The SLE plasma thermograms also offer exciting new avenues of research, with the binary classification of SLE against non-SLE alternatives representing only a starting point for the potential of the diagnostic technique. The goal of the author is to continue a theme of applied statistical research to interdisciplinary problems, with FDA, classification, and ensemble learning driving the investigations.

FDA has the potential to be a flexible tool with many extensions for interdisciplinary study. The ESFuNC algorithm demonstrates the potential of FDA classification and the need to develop novel approaches to using functional covariates. The ESFuNC algorithm represents an initial approach to the combination of multivariate functional data, as segmentation was restricted to equal partitions. How to loosen these restrictions and produce final ensembles containing unequal segmented-FDOs is expected to be related with additional computational complexity. PPEs are an initial step into evaluating how mixtures of large sets of classifiers can be completed. LR estimation of ensembles within the ESFuNC algorithm has also yet to be investigated, with LR

estimation of PPEs showing promising potential. Predictor grouping estimates from fused-LASSO could also impact the ESFuNC algorithm. If it is possible to estimate which regions of the predictor domain should be grouped, these regions could be used to influence the selection of the intervals defining the segmented-FDOs.

Functional and shape analysis also provides many unique avenues for future investigation. Functional representations of SLE and non-SLE patients were used to produce the unique shapes provided in Figure 19. The shapes are produced by plotting amplitude values for first and second derivative approximations against one another. The figure shows the mean shapes produced from SLE and non-SLE alternatives; there are several clear distinctions in the shapes produced. Shape analysis of such images could be an effective methodology for population discrimination that has yet to be investigated (Srivastava and Klassen 2016). This type of analysis also defines a system of dynamic equations for the description of the shape parameters. By evaluating the relationship between first and second derivative amplitudes, it may be possible to design a set of differential equations capable of describing the difference between SLE and non-SLE signatures. Statistical classification based on dynamical systems is a potential field of future study, with FDA providing a foundation for developing such classifiers.

The SLE plasma thermogram data set will also be exploited for additional clinical challenges. One major study of interest will investigate if additional morbidities can be identified using only plasma thermogram information. Within the SLE thermogram data set are additional partitions beyond SLE vs. non-SLE alternatives. Other autoimmune disorders, such as arthritis, fibromyalgia, and scleroderma are also represented within the

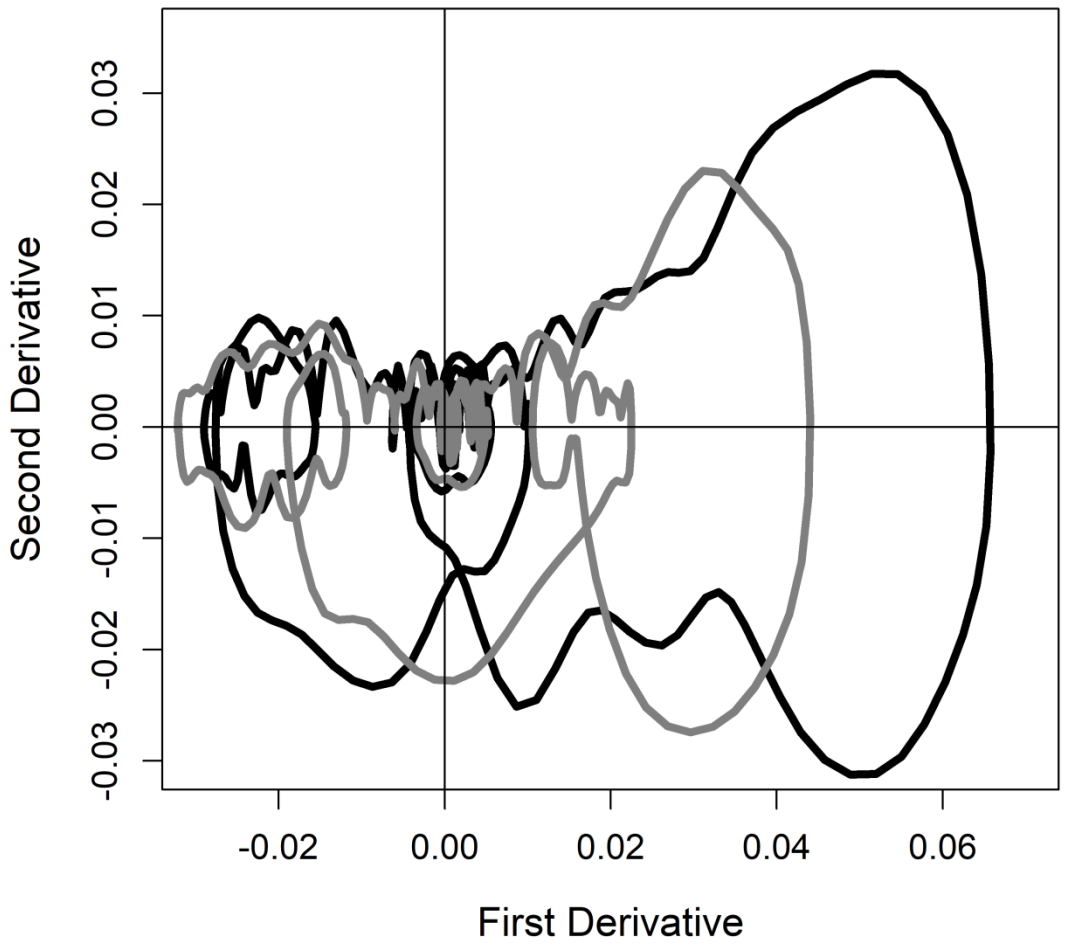

Figure 19. Shapes produced by plotting first derivative against second derivative amplitudes from functional approximations to the SLE plasma thermograms. The mean shape produced from non-SLE curves (black) is overlaid with the mean shape from SLE curves (grey). The functions correspond to the GCV basis with additional smoothing using a penalty of 0.01 to ensure a smoothed second derivative approximation.

data set. Hence, a challenging clinical application would be the identification of patients suffering from co-morbidities. Developing classifiers for evaluating such cases will be a step further than the binary applications within this dissertation. Classifiers for identification of co-morbidities will be greatly aided by the use of additional predictor information, including patient information as well as serological and immunological markers.

REFERENCES

Aguilera, A. M., Escabias, M., & Valderrama, M. J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, *50*(8), 1905-1924.

Aguilera, A. M., Escabias, M., Valderrama, M. J., & Aguilera-Morillo, M. C. (2013). Functional analysis of chemometric data. *Open Journal of Statistics*, *3*(05), 334.

Analytics, R., & Weston, S. (2014). doParallel: Foreach parallel adaptor for the parallel package. *R package version*, *1.4.2*.

Analytics, R., & Weston, S. (2014b). foreach: Foreach looping construct for R. *R package version*, *1.4.2*.

Baillo, A., & Cuevas, A. (2008). Supervised functional classification: a theoretical remark and some comparisons. *arXiv preprint arXiv:0806.2831*.

Berrendero, J. R., Cuevas, A., & Torrecilla, J. L. (2016). Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica*, 619-638.

Buscaglia, R., Gray, R. D., & Chaires, J. B. (2013). Thermodynamic characterization of human telomere quadruplex unfolding. *Biopolymers*, *99*(12), 1006-1018.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004, July). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning* (p. 18). ACM.

Cashman, D. J., Buscaglia, R., Freyer, M. W., Dettler, J., Hurley, L. H., & Lewis, E. A. (2008). Molecular modeling and biophysical analysis of the c-MYC NHE-III 1 silencer element. *Journal of molecular modeling*, *14*(2), 93-101.

Chaires, J. B., Trent, J. O., Gray, R. D., Dean, W. L., Buscaglia, R., Thomas, S. D., & Miller, D. M. (2014). An improved model for the hTERT promoter quadruplex. *PloS one*, *9*(12), e115580.

Chan, K. S., & Chen, K. (2011). Subset ARMA selection via the adaptive Lasso. *Statistics and its Interface*, *4*(2), 197-205.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391-403.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf*.

Daly, R., Partovi, R., & Davidson, P. (2017, October). Lupus Diagnosis: Process and Patient Experience. In *ARTHRITIS & RHEUMATOLOGY* (Vol. 69). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.

Delaigle, A., Hall, P., & Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika*, *99*(2), 299-313.

Dettler, J. M., Buscaglia, R., Cui, J., Cashman, D., Blynn, M., & Lewis, E. A. (2010). Biophysical characterization of an ensemble of intramolecular i-motifs formed by the human c-MYC NHE III1 P1 promoter mutant sequence. *Biophysical journal*, *99*(2), 561-567.

Dettler, J. M., Buscaglia, R., Le, V. H., & Lewis, E. A. (2011). DSC deconvolution of the structural complexity of c-MYC P1 promoter G-quadruplexes. *Biophysical journal*, *100*(6), 1517-1525.

Dietterich, T. G. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, *2*, 110-125.

Dony, R. (2001). The Transform and Data Compression Handbook. *Karhunen-Loève Transform. CRC Press, Boca Raton*.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 325-327.

Escabias, M., Aguilera, A. M., & Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, *16*(3-4), 365-384.

Febrero-Bande, M., & de la Fuente, M. O. (2012). Statistical computing in functional data analysis: The R package fda. usc. *Journal of Statistical Software*, *51*(4), 1-28.

Febrero-Bande, M., & González-Manteiga, W. (2013). Generalized additive models for functional data. *Test*, *22*(2), 278-292.

Ferraty, F., & Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, *44*(1-2), 161-173.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Fish, D. J., Brewood, G. P., Kim, J. S., Garbett, N. C., Chaires, J. B., & Benight, A. S. (2010). Statistical analysis of plasma thermograms measured by differential scanning calorimetry. *Biophysical chemistry*, *152*(1-3), 184-190.

Fisher, R. A. (1925, July). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 22, No. 5, pp. 700-725). Cambridge University Press.

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

Freyer, M. W., Buscaglia, R., Kaplan, K., Cashman, D., Hurley, L. H., & Lewis, E. A. (2007). Biophysical studies of the c-MYC NHE III1 promoter: model quadruplex interactions with a cationic porphyrin. *Biophysical journal*, *92*(6).

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, pp. 337-387). New York: Springer series in statistics.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, *1*(4).

Garbett, N. C., & Brock, G. N. (2016). Differential scanning calorimetry as a complementary diagnostic tool for the evaluation of biological samples. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1860*(5), 981-989.

Garbett, N. C., Brock, G. N., Chaires, J. B., Mekmaysy, C. S., DeLeeuw, L., Sivils, K. L., ... & Jarjour, W. N. (2017). Characterization and classification of lupus patients based on plasma thermograms. *PloS one*, *12*(11), e0186398.

Garbett, N. C., Mekmaysy, C. S., DeLeeuw, L., & Chaires, J. B. (2015). Clinical application of plasma thermograms. Utility, practical approaches and considerations. *Methods*, *76*, 41-50.

Garbett, N. C., Mekmaysy, C. S., Helm, C. W., Jenson, A. B., & Chaires, J. B. (2009). Differential scanning calorimetry of blood plasma for clinical diagnosis and monitoring. *Experimental and Molecular Pathology*, *86*(3), 186-191.

Garbett, N. C., Merchant, M. L., Chaires, J. B., & Klein, J. B. (2013). Calorimetric analysis of the plasma proteome: Identification of type 1 diabetes patients with early renal function decline. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1830*(10), 4675-4680.

Garbett, N. C., Merchant, M. L., Helm, C. W., Jenson, A. B., Klein, J. B., & Chaires, J. B. (2014). Detection of cervical cancer biomarker patterns in blood plasma and urine by differential scanning calorimetry and mass spectrometry. *PloS one*, *9*(1), e84710.

Garbett, N. C., Miller, J. J., Jenson, A. B., & Chaires, J. B. (2007). Calorimetric analysis of the plasma proteome. In *Seminars in nephrology* (Vol. 27, No. 6, pp. 621-626). Elsevier.

Garbett, N. C., Miller, J. J., Jenson, A. B., & Chaires, J. B. (2008). Calorimetry outside the box: a new window into the plasma proteome. *Biophysical journal*, *94*(4), 1377-1383.

Garbett, N. C., Miller, J. J., Jenson, A. B., Miller, D. M., & Chaires, J. B. (2007). Interrogation of the plasma proteome with differential scanning calorimetry. *Clinical chemistry*, *53*(11), 2012-2014.

Goeman, J., Meijer, R., & Chaturvedi, N. (2012). penalized: L1 (lasso and fused lasso) and L2 (ridge) Penalized Estimation in GLMs and in the Cox Model. *URL http://cran. r-project. org/web/packages/penalized/index. html*.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical journal*, *52*(1), 70-84.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215-223.

Górecki, T., & Krzyśko, M. (2012). Functional principal components analysis. *Data Analysis Methods and its Applications, CH Beck, Warszawa*, 71-87.

Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN 98* (pp. 201-206). Springer, London.

Grandvalet, Y., & Canu, S. (1999). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *Advances in neural information processing systems* (pp. 445-451).

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W., & Lausen, B. (2016). Ensemble of a subset of kNN classifiers. *Advances in Data Analysis and Classification*, 1-14.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 73-102.

Hastie, T. J., & Tibshirani, R. J. (1990). Generalized additive models, volume 43 of Monographs on Statistics and Applied Probability.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Overview of supervised learning. In *The elements of statistical learning* (pp. 9-40). Springer, New York, NY.

Hechenbichler, K., & Schliep, K. (2004). Weighted k-nearest-neighbor techniques and ordinal classification.

Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE transactions on pattern analysis and machine intelligence*, *16*(1), 66-75.

Hochberg, M. C. (1997). Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis & Rheumatology*, *40*(9), 1725-1725.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Höhne, G. W. H., Hemminger, W., & Flammersheim, H. J. (1996). Theoretical fundamentals of differential scanning calorimeters. In *Differential Scanning Calorimetry* (pp. 21-40). Springer, Berlin, Heidelberg.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

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

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Krier, C., François, D., Rossi, F., & Verleysen, M. (2009, April). Supervised variable clustering for classification of NIR spectra. In *ESANN*.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... & Kenkel, B. (2015). caret: Classification and regression training. R package version 6.0–21. *CRAN, Vienna, Austria*.

Lee, S. H., Yu, D., Bachman, A. H., Lim, J., & Ardekani, B. A. (2014). Application of fused lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease. *Journal of neuroscience methods*, *221*, 78-84.

Li, B., & Yu, Q. (2008). Classification of functional data: A segmentation approach. *Computational Statistics & Data Analysis*, *52*(10), 4790-4800.

Liu, J., Yuan, L., & Ye, J. (2010, July). An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 323-332). ACM.

Liu, Y., & Yao, X. (1999). Ensemble learning via negative correlation. *Neural networks*, *12*(10), 1399-1404.

Müller, H. G., & Stadtmüller, U. (2005). Generalized functional linear models. *the Annals of Statistics*, *33*(2), 774-805.

Müller, H. G., & Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, *103*(484), 1534-1544.

Muller, K. R., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, *12*(2), 181-201.

Nagesh, N., Buscaglia, R., Dettler, J. M., & Lewis, E. A. (2010). Studies on the site and mode of TMPyP4 interactions with Bcl-2 promoter sequence G-Quadruplexes. *Biophysical journal*, *98*(11), 2628-2633.

Narain, S., Richards, H. B., Satoh, M., Sarmiento, M., Davidson, R., Shuster, J., ... & Reeves, W. H. (2004). Diagnostic accuracy for lupus and other systemic autoimmune diseases in the community setting. *Archives of internal medicine*, *164*(22), 2435-2441.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065-1076.

Porro-Muñoz, D., Talavera, I., Duin, R. P., Hernández, N., & Orozco-Alzate, M. (2011). Dissimilarity representation on functional spectral data for classification. *Journal of Chemometrics*, *25*(9), 476-486.

Rai, S. N., Pan, J., Cambon, A., Chaires, J. B., & Garbett, N. C. (2013). Group classification based on high-dimensional data: application to differential scanning calorimetry plasma thermogram analysis of cervical cancer and control samples. *Open Access Medical Statistics*, *3*, 1-9.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). Functional Data Analysis with R and MATLAB. Use R, 1-207.

Ramsay, J. O., & Silverman, B. W. (2005). Springer series in statistics. In *Functional data analysis*. Springer.

Ramsay, J. O., Wickham, H., Graves, S., & Hooker, G. (2014) fda: Functional Data Analysis. *R package version 2.4.4.*

Ramsay, J. O. (2006). *Functional data analysis*. John Wiley & Sons, Inc..

Ramsay, J. O., & Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Ratcliffe, S. J., Heller, G. Z., & Leader, L. R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in medicine*, *21*(8), 1115-1127.

Rizwan, M., & Anderson, D. V. Comparison of Distance Metrics for Phoneme Classification based on Deep Neural Network Features and Weighted k-NN Classifier.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1-39.

Rosen, B. E. (1996). Ensemble learning using decorrelated neural networks. *Connection science*, *8*(3-4), 373-384.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517.

Saldana, D. F., & Feng, Y. (2016). SIS: An r package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software*.

Sturtevant, J. M. (1987). Biochemical applications of differential scanning calorimetry. *Annual review of physical chemistry*, *38*(1), 463-488.

Thodberg, H. H. (1995). Tecator data set. *Contained in StatLib Datasets Archive*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273-282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91-108.

Todinova, S., Krumova, S., Kurtev, P., Dimitrov, V., Djongov, L., Dudunkov, Z., & Taneva, S. G. (2012). Calorimetry-based profiling of blood plasma from colorectal cancer patients. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1820*(12), 1879-1885.

Verleysen, M., & François, D. (2005, June). The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks* (pp. 758-770). Springer, Berlin, Heidelberg.

Wang, H., & Leng, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, *102*(479), 1039-1048.

Yang, Y., Zou, H., Yang, M. Y., & Matrix, D. (2017). Package ‘gcdnet’.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, *5*(Oct), 1205-1224.

Zapf, I., Fekecs, T., Ferencz, A., Tizedes, G., Pavlovics, G., Kálmán, E., & Lőrinczy, D. (2011). DSC analysis of human plasma in breast cancer patients. *Thermochimica acta*, *524*(1-2), 88-91.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, *37*(4), 1733.

APPENDIX A

SUPPLEMENTAL TABLES AND FIGURES

Original + First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.821 (0.051)	0.826 (0.071)	0.816 (0.070)
RIDGE	0.866 (0.043)	0.874 (0.061)	0.858 (0.060)
ENET	0.861 (0.042)	0.865 (0.060)	0.856 (0.059)
adap-ENET	0.859 (0.045)	0.870 (0.060)	0.847 (0.066)
LASSO	0.861 (0.043)	0.867 (0.060)	0.855 (0.060)
adap-LASSO	0.859 (0.045)	0.867 (0.061)	0.851 (0.064)
LDA	0.854 (0.047)	**0.879** (0.059)	0.827 (0.072)
QDA	**0.879** (0.037)	0.833 (0.064)	**0.925** (0.053)
KNN	0.821 (0.048)	0.823 (0.073)	0.818 (0.066)
Original + First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.807 (0.051)	0.811 (0.072)	0.803 (0.076)
RIDGE	0.893 (0.038)	0.894 (0.053)	**0.892** (0.054)
ENET	**0.888** (0.041)	0.885 (0.058)	0.891 (0.055)
adap-ENET	0.867 (0.044)	0.875 (0.058)	0.858 (0.058)
LASSO	0.882 (0.043)	0.883 (0.060)	0.881 (0.059)
adap-LASSO	0.863 (0.045)	0.870 (0.063)	0.856 (0.060)
LDA	0.847 (0.043)	0.858 (0.066)	0.837 (0.065)
QDA	DNC	DNC	DNC
KNN	0.881 (0.039)	**0.931** (0.044)	0.830 (0.073)

Table A1. SLE TRUNC combined predictor classification performance of the nine classifiers as given by the accuracy, specificity, and sensitivity. Predictors were produced by combining discretized samples from original curves with first derivative (Original + First Derivative) or with both first and second derivative approximations (Original + First Derivative + Second Derivative). The test set mean and standard deviation for each metric is recorded. DNC represents solutions that did not converge.

Original + First Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.806 (0.057)	0.827 (0.077)	0.785 (0.097)
RIDGE	0.864 (0.041)	0.864 (0.059)	**0.864** (0.063)
ENET	**0.868** (0.041)	**0.876** (0.056)	0.860 (0.063)
adap-ENET	0.861 (0.048)	0.872 (0.060)	0.850 (0.073)
LASSO	0.863 (0.041)	0.871 (0.057)	0.855 (0.064)
adap-LASSO	0.854 (0.045)	0.864 (0.060)	0.843 (0.067)
LDA	0.853 (0.042)	0.867 (0.060)	0.838 (0.065)
QDA	DNC	DNC	DNC
KNN	0.775 (0.053)	0.746 (0.079)	0.804 (0.077)
Original + First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
LR	0.792 (0.058)	0.836 (0.091)	0.748 (0.128)
RIDGE	**0.878** (0.041)	**0.881** (0.055)	**0.874** (0.062)
ENET	0.873 (0.040)	0.879 (0.056)	0.866 (0.061)
adap-ENET	0.871 (0.043)	0.880 (0.058)	0.861 (0.064)
LASSO	0.875 (0.041)	0.878 (0.057)	0.872 (0.062)
adap-LASSO	0.866 (0.042)	0.874 (0.055)	0.858 (0.065)
LDA	0.853 (0.042)	0.867 (0.060)	0.838 (0.065)
QDA	DNC	DNC	DNC
KNN	0.849 (0.045)	0.828 (0.069)	0.871 (0.067)

Table A2. GCV TRUNC combined predictor classification performance of the nine classifiers as given by the accuracy, specificity, and sensitivity. Predictors were produced by combining discretized samples from original curves with first derivative (Original + First Derivative) or both first and second derivative approximations (Original + First Derivative + Second Derivative). The test set mean and standard deviation for each metric is recorded. DNC represents solutions that did not converge.

Original + First Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.756 (0.055)	**0.743** (0.081)	0.769 (0.078)
FGSAM	**0.762** (0.055)	0.737 (0.080)	**0.787** (0.076)
FGKAM	0.719 (0.057)	0.691 (0.086)	0.748 (0.083)
Original + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	0.775 (0.053)	**0.787** (0.077)	0.763 (0.077)
FGSAM	**0.784** (0.052)	0.782 (0.074)	0.785 (0.072)
FGKAM	0.743 (0.056)	0.698 (0.084)	**0.788** (0.073)
First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	**0.805** (0.051)	**0.793** (0.076)	**0.816** (0.073)
FGSAM	0.793 (0.048)	0.785 (0.073)	0.800 (0.071)
FGKAM	0.749 (0.055)	0.740 (0.082)	0.759 (0.079)
Original + First Derivative + Second Derivative			
Method	Accuracy	Sensitivity	Specificity
FGLM	**0.786** (0.052)	0.771 (0.079)	**0.801** (0.076)
FGSAM	0.784 (0.051)	**0.778** (0.078)	0.791 (0.073)
FGKAM	0.743 (0.056)	0.711 (0.084)	0.775 (0.076)

Table A3. FLR classification results using multiple functional covariates based on the GCV FDO. Performance is summarized by accuracy, specificity, and sensitivity. Given is the mean and standard deviation for each metric.

Naïve Ensemble			
Method	Accuracy	Sensitivity	Specificity
LR	0.816 (0.046)	0.833 (0.070)	0.798 (0.079)
RIDGE	0.882 (0.041)	0.874 (0.057)	0.890 (0.058)
ENET	0.878 (0.042)	0.879 (0.058)	0.878 (0.061)
LASSO	0.878 (0.040)	**0.881** (0.057)	0.874 (0.060)
LDA	0.855 (0.043)	0.867 (0.061)	0.842 (0.063)
QDA*	**0.901** (0.035)	0.878 (0.058)	**0.924** (0.049)
KNN	0.826 (0.051)	0.815 (0.070)	0.838 (0.069)
FGLM	0.740 (0.059)	0.720 (0.084)	0.762 (0.081)
FGSAM	0.771 (0.054)	0.747 (0.075)	0.795 (0.076)
FGKAM	0.738 (0.055)	0.716 (0.083)	0.761 (0.080)
FKNN	0.817 (0.050)	0.808 (0.071)	0.827 (0.071)

Table A4. Performance of naïve ensembles summarized by accuracy, sensitivity, and specificity for all classifiers based on the GCV FDO. Naïve ensembles were produced through estimated classes using original curves along with first and second derivatives. Given is the mean and standard deviation for each metric. *Results were obtained using GCV TRUNC predictor set.

Equally Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.821 (0.046)	0.821 (0.045)	0.814 (0.047)	0.816 (0.046)
RIDGE	0.868 (0.043)	0.878 (0.041)	0.879 (0.041)	0.882 (0.041)
ENET	0.874 (0.040)	0.885 (0.040)	0.879 (0.041)	0.882 (0.041)
LASSO	0.873 (0.040)	0.883 (0.040)	0.879 (0.040)	0.883 (0.041)
LDA	0.854 (0.043)	0.856 (0.043)	0.856 (0.045)	0.855 (0.044)
QDA*	**0.905** (0.035)	0.899 (0.036)	**0.906** (0.033)	**0.901** (0.034)
KNN	0.782 (0.054)	0.898 (0.041)	0.888 (0.041)	0.873 (0.046)
FGLM	0.745 (0.058)	0.734 (0.054)	0.747 (0.056)	0.748 (0.058)
FGSAM	0.767 (0.053)	0.764 (0.054)	0.771 (0.054)	0.774 (0.052)
FGKAM	0.717 (0.058)	0.744 (0.056)	0.747 (0.057)	0.732 (0.057)
FKNN	0.778 (0.055)	**0.903** (0.041)	0.897 (0.040)	0.881 (0.044)
Accuracy Weighted				
Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
LR	0.798 (0.053)	0.796 (0.052)	0.786 (0.053)	0.816 (0.046)
RIDGE	0.867 (0.043)	0.877 (0.040)	0.879 (0.041)	0.882 (0.041)
ENET	0.873 (0.040)	0.883 (0.040)	0.879 (0.041)	0.881 (0.041)
LASSO	0.873 (0.040)	0.882 (0.040)	0.878 (0.040)	0.882 (0.042)
LDA	0.854 (0.043)	0.856 (0.043)	0.856 (0.045)	0.855 (0.044)
QDA*	**0.901** (0.036)	0.894 (0.036)	**0.901** (0.036)	**0.901** (0.034)
KNN	0.782 (0.054)	0.894 (0.041)	0.883 (0.041)	0.867 (0.046)
FGLM	0.747 (0.058)	0.735 (0.055)	0.749 (0.057)	0.749 (0.057)
FGSAM	0.768 (0.052)	0.765 (0.054)	0.772 (0.054)	0.775 (0.052)
FGKAM	0.717 (0.058)	0.745 (0.054)	0.750 (0.056)	0.734 (0.056)
FKNN	0.777 (0.055)	**0.904** (0.043)	0.899 (0.041)	0.887 (0.041)

Table A5. Weighted ensemble results for all classifiers based on the GCV FDO. Ensemble accuracies for all combinations of the three classifiers (D^0 : original curve, D^1 : first derivative, and D^2 : second derivative) are given. Performance is summarized by accuracy with the test set mean and standard deviation recorded. *Results were obtained using GCV TRUNC predictor set.

Greedy Ensemble Strategy	
Classifier	Optimal Segments Included
KNN	$F^{(1)}_{15,1,6,7,9-13,15}$, $F^{(2)}_{17,2}$
Tri-WKNN	$F^{(1)}_{16,2-11,13,15}$
Norm-WKNN	$F^{(1)}_{29,2,3,7,8,12,15-20,22,24,27,29}$
Tri-PW	$F^{(0)}_{24,1,3-6,8,12,15,17-19,21,23-24}$
Norm-PW	$F^{(0)}_{17,1,2,5,11,13,14,17}$, $F^{(1)}_{17,1,5,9-11,13-17}$
Combined Ensemble Strategy	
Classifier	Optimal Segments Included
KNN	$F^{(0)}_{5,1,2,4}$, $F^{(1)}_{5,1,3-5}$
Tri-WKNN	$F^{(0)}_{5,3,5}$, $F^{(1)}_{5,3-5}$, $F^{(2)}_{5,2}$
Norm-WKNN	$F^{(0)}_{5,1,2,4}$, $F^{(1)}_{5,1,3-5}$
Tri-PW	$F^{(0)}_{7,3,6}$, $F^{(1)}_{7,1,4-6}$, $F^{(2)}_{7,1,3,7}$
Norm-PW	$F^{(0)}_{5,3-5}$, $F^{(1)}_{5,1-5}$, $F^{(2)}_{5,5}$
Hierarchical Ensemble Strategy	
Classifier	Optimal Segments Included
KNN	$F^{(0)}_{7,4}$, $F^{(1)}_{6,2-5}$, $F^{(2)}_{5,2}$
Tri-WKNN	$F^{(0)}_{14,4,8,10,14}$, $F^{(1)}_{6,1-5}$, $F^{(2)}_{5,5}$
Norm-WKNN	$F^{(0)}_{17,5,9,17}$, $F^{(1)}_{10,1,4,6-9}$, $F^{(2)}_{1,1}$
Tri-PW	$F^{(0)}_{3,2}$, $F^{(1)}_{8,1,3,5,6,8}$, $F^{(2)}_{6,1,6}$
Norm-PW	$F^{(0)}_{6,3-5}$, $F^{(1)}_{6,1,2,4-6}$, $F^{(2)}_{3,1,3}$

Table A6. Optimized segmented-FDOs included in ESFuNC final ensemble for the SLE plasma thermogram data set. Segmented-FDOs are given with superscripted derivative order. Subscript gives the total segmentation size followed by the segmented-FDOs included in the final ensemble.

Greedy Ensemble Strategy	
Classifier	Optimal Segments Included
KNN	$F_{4,2}^{(2)}$
Tri-WKNN	$F_{2,1}^{(1)}$, $F_{6,1,3}^{(2)}$
Norm-WKNN	$F_{2,1}^{(1)}$, $F_{6,3,4}^{(2)}$
Tri-PW	$F_{2,1}^{(0)}$, $F_{13,5}^{(1)}$
Norm-PW	$F_{5,3}^{(1)}$
Combined Ensemble Strategy	
Classifier	Optimal Segments Included
KNN	$F_{4,4}^{(0)}$, $F_{4,2}^{(2)}$
Tri-WKNN	$F_{4,4}^{(0)}$, $F_{4,4}^{(1)}$, $F_{4,2}^{(2)}$
Norm-WKNN	$F_{4,2}^{(1)}$, $F_{4,1,2}^{(2)}$
Tri-PW	$F_{6,2,3}^{(2)}$
Norm-PW	$F_{2,1}^{(1)}$, $F_{2,1}^{(2)}$
Hierarchical Ensemble Strategy	
Classifier	Optimal Segments Included
KNN	$F_{4,2}^{(2)}$
Tri-WKNN	$F_{4,2}^{(2)}$
Norm-WKNN	$F_{4,2}^{(2)}$
Tri-PW	$F_{2,1}^{(1)}$, $F_{3,1}^{(2)}$
Norm-PW	$F_{2,1}^{(1)}$, $F_{2,1}^{(2)}$

Table A7. Optimized segmented-FDOs included in ESFuNC final ensemble for the Tecator data set. Segmented-FDOs are given with superscripted derivative order. Subscript gives the total segmentation size followed by the segmented-FDOs included in the final ensemble.

Greedy Ensemble Strategy	
Method	Optimal Segments Included
KNN	$F_{2,1,2}^{(0)}$, $F_{4,1,2,4}^{(1)}$, $F_{4,1,3,4}^{(2)}$
Tri-WKNN	$F_{2,1,2}^{(0)}$, $F_{4,1,3}^{(1)}$, $F_{4,1-4}^{(2)}$
Norm-WKNN	$F_{2,1}^{(0)}$, $F_{4,1-4}^{(1)}$, $F_{4,3,4}^{(2)}$
Tri-PW	$F_{3,1}^{(0)}$, $F_{10,1,3-5}^{(1)}$
Norm-PW	$F_{2,1}^{(0)}$, $F_{8,1,2,4,5}^{(1)}$, $F_{8,2,3,5,6}^{(2)}$
Combined Ensemble Strategy	
Method	Optimal Segments Included
KNN	$F_{4,1,2,4}^{(0)}$, $F_{4,1-4}^{(1)}$, $F_{4,1,4}^{(2)}$
Tri-WKNN	$F_{6,1,2,4,6}^{(0)}$, $F_{6,1-3,6}^{(1)}$, $F_{6,5}^{(2)}$
Norm-WKNN	$F_{2,1}^{(0)}$, $F_{2,1,2}^{(1)}$, $F_{2,2}^{(2)}$
Tri-PW	$F_{3,1-3}^{(0)}3$, $F_{3,1-3}^{(1)}$, $F_{3,1}^{(2)}$
Norm-PW	$F_{2,1,2}^{(0)}$, $F_{2,1,2}^{(1)}$, $F_{2,1}^{(2)}$
Hierarchical Ensemble Strategy	
Method	Optimal Segments Included
KNN	$F_{2,1}^{(0)}$, $F_{12,1,2,6,9,12}^{(1)}$, $F_{9,8}^{(2)}$
Tri-WKNN	$F_{2,1}^{(0)}$, $F_{3,1,3}^{(1)}$, $F_{3,2}^{(2)}$
Norm-WKNN	$F_{2,1}^{(0)}$, $F_{18,1,4,9,14,17}^{(1)}$, $F_{13,2,7,8,10}^{(2)}$
Tri-PW	$F_{3,1}^{(0)}$, $F_{11,1,2,5,11}^{(1)}$, $F_{14,6,7}^{(2)}$
Norm-PW	$F_{2,1}^{(0)}$, $F_{3,1,3}^{(1)}$, $F_{8,3,6}^{(2)}$

Table A8. Optimized segmented-FDOs included in ESFuNC final ensemble for the Phoneme data set. Segmented-FDOs are given with superscripted derivative order. Subscript gives the total segmentation size followed by the segmented-FDOs included in the final ensemble.

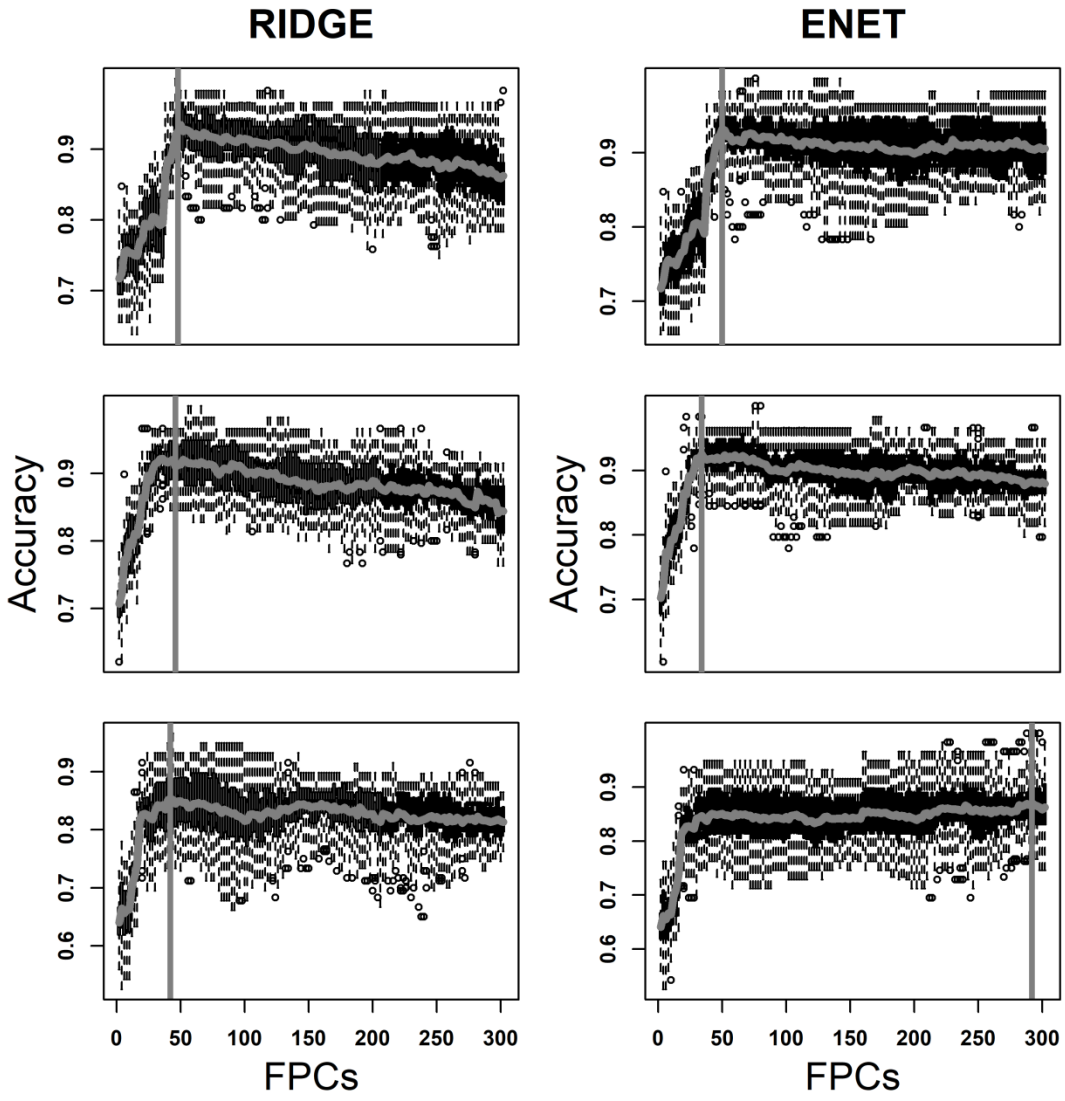

Figure A1. Summary of FPCA-based RIDGE and ENET classification performance using an increasing number of FPCs. Results are shown based on FPCA of the SLE FDO original curves and its first and second derivatives. Original curves are shown on top, followed by first derivatives in the middle and second derivatives on the bottom. Boxplots of classification accuracies resulting from KCV are given for each value of k. Grey line indicates the mean test set accuracy at each FPC. The vertical grey line represents the number of FPCs that returns the highest mean test set accuracy.

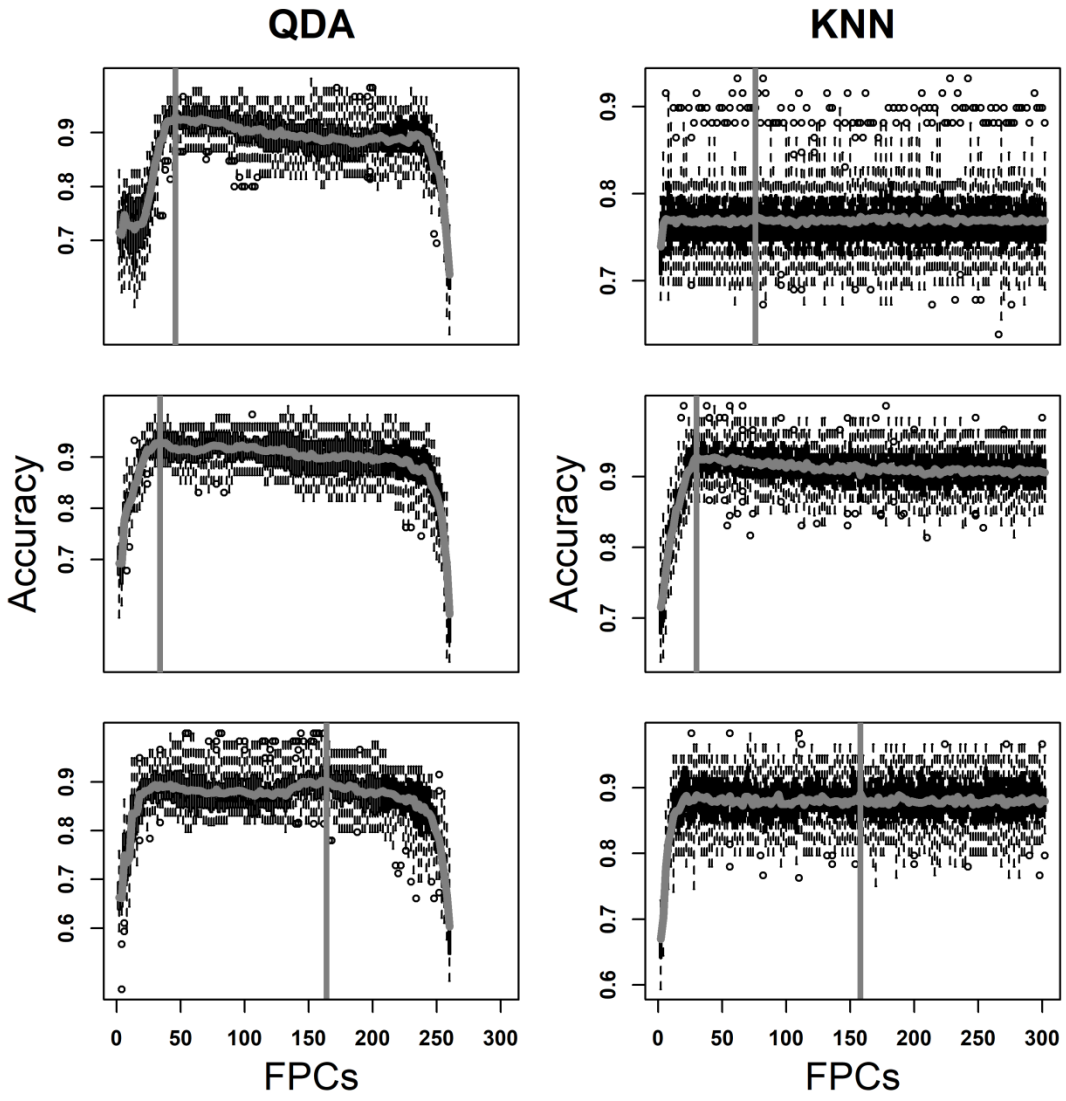

Figure A2. Summary of FPCA-based QDA and KNN classification performance using an increasing number of FPCs. Results are shown based on FPCA of the SLE FDO original curves and its first and second derivatives. Original curves are shown on top, followed by first derivatives in the middle and second derivatives on the bottom. Boxplots of classification accuracies resulting from KCV are given for each value of k. Grey line indicates the mean test set accuracy at each FPC. The vertical grey line represents the number of FPCs that returns the highest mean test set accuracy.

Method	$D^0 + D^1$	$D^0 + D^2$	$D^1 + D^2$	$D^0 + D^1 + D^2$
RIDGE	0.851 (0.044)	0.867 (0.042)	0.890 (0.042)	0.888 (0.032)
ENET	**0.887** (0.039)	**0.903** (0.039)	**0.905** (0.039)	**0.897** (0.041)
LASSO	0.882 (0.041)	0.885 (0.040)	0.897 (0.040)	0.893 (0.040)

Table A9. LR-estimated PPEs using combined PPC predictor sets. LR was performed on predictors combined from each combination of original (D^0), first derivative (D^1), and second derivative (D^2) PPCs. Test set accuracy mean and standard deviation is reported for each combination.