Computing Obesity: Signal Processing and Machine Learning Applied to Predictive Modeling of Clinical Weight-Loss

by

Craig A. Biwer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2017

Doctoral Committee:

Associate Professor Kayvan Najarian, Chair
Professor Charles F. Burant
Professor Harm Derksen
Professor H.V. Jagadish
Professor Gilbert S. Omenn

Craig A. Biwer

cbiwer@umich.edu

ORCID id: 0000-0001-9508-1111

To my parents, Kelli, and Cat

ACKNOWLEDGEMENTS

Primary thanks go to Dr. Kayvan Najarian for mentoring me these past four years. He always had time to help, and his boundless optimism was matched only by his patience.

Additional thanks to the rest of my committee: Dr. Charles F. Burant, Dr. Harm Derksen, Dr. H.V. Jagadish, and Dr. Gilbert Omenn. Their guidance and advice have been invaluable in this work.

Thank you to Dr. Amy Rothberg and Dr. Heidi IglayReger for helping with the data, and to Julia Eussen for answering my endless questions. Further thanks to all of the people in the Department of Computational Medicine and Bioinformatics; the Michigan Center for Integrative Research in Critical Care; the Department of Metabolism, Endocrinology & Diabetes; the Michigan Center for Diabetes Translational Research; the Blue Care Network; and the Blue Cross Blue Shield of Michigan Foundation.

Finally, my eternal gratitude to Kelli and my family. I could not have done this without your support, encouragement, and love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

AUC Area Under the ROC Curve

BMI Body Mass Index

LCG Laplacian of Correlation Graph classification

MWMP University of Michigan Weight Management Program clinic

RMR Resting Metabolic Rate

ROC Receiver Operating Characteristic

SVM Support Vector Machine

ABSTRACT

Overweight and obesity are highly prevalent in the United States, with over two-thirds of the adult population classified as overweight and over one-third as obese. Associated with a number of serious diseases, these conditions have been shown to increase the risk of issues such as hypertension, type 2 diabetes mellitus, and depression, among others. All told, overweight and obesity place a significant burden on the modern healthcare industry, with estimates on the cost as high as $210 billion per year. Many obese individuals attempt to lose weight but, following a loss, a series of neurobehavioral mechanisms activate that commonly result in weight regain. There are no methods to date for determining *a priori* who will successfully lose weight and maintain the loss, nor have any definitive factors been identified that can be used to predict who will see a long-term reduction in his or her weight-related medication regimen.

These problems, along with many others in the clinical field, stand to benefit from the application of signal processing and machine learning methods. To begin addressing these issues, participants in a two-year weight-loss study are split into two groups based on their ability to lose weight while dieting and to maintain at least a portion of that loss. Utilizing accelerometer data collected before each subject's diet, a windowed approach to persistent homology is used to show a clear difference in the intra-group similarities between the movement profiles of the two groups ($p = 1.505 \times 10^{-23}$). This application of persistent homology presents a novel take on the topological method, allowing for more clinically relevant results by placing limits

on the time frame in which two activities can be considered related. By expanding upon and investigating the measured difference, insights can be gained on how movement affects diet efficacy. From the same study, a separate metric for success based on an individual's medication history is developed. Using features extracted from physiological signals collected both before and after the diet, a Naïve Bayes model is generated. After reducing the feature set to filter out noise, this model is shown to be able to predict, with an accuracy of nearly 86%, which individuals will require more prescription medications and which will require fewer a year and a half later. This indicates that weight loss can have a lasting impact on health, regardless of future weight regain, and has major implications for the pharmacological industry. Furthering these results, a new machine learning algorithm is developed and presented. Meant for noisy datasets, Laplacian of Correlation Graph Classification shows improvements in accuracy and robustness over standard machine learning algorithms when applied to unreduced feature sets of varying sizes. This method not only removes the risk of excluding potentially useful data through feature selection, but it can also provide clinically relevant insights into the underlying relationships between disparate measurements.

CHAPTER I

Introduction

1.1 Problem Statement

With the prevalence of overweight and obesity on the rise, the need to develop computational approaches to predict which individuals will benefit from dieting is also increasing. While weight loss is an obvious indication of success, it must be balanced against any future weight regain. Similarly, a reduction in an individual's weight-related medication regimen may be an indication of an improvement of overall health, but it must be maintained in order to be considered a long-term improvement. In addition to the metric used to determine success, the means by which the data are gathered must also be considered. While in-lab tests may yield more reliable and detailed information, at-home monitoring involves significantly less time and effort, especially for the clinicians. This thesis begins to address the issue of diet-induced weight loss through the application of signal processing and machine learning algorithms: by studying available physiological data, recorded both in controlled laboratory settings and at home, patterns begin to emerge. Through the use of methods introduced here, it becomes possible to make predictions of the long-term efficacy of dieting for any given individual. This has major implications not only in the clinical field but in the healthcare industry as a whole.

A background on obesity, as well as an overview of signal processing and machine

learning, is found in the next section. In Chapter II an application of persistent homology to recorded movement data is presented. Chapter III demonstrates the successful application of machine learning algorithms to physiological data, where the labels are defined by a reduction in medication regimen. This is followed by the introduction of a graph-based machine learning algorithm in Chapter IV that not only allows for the integration of data from disparate sources but also performs better than standard approaches on unfiltered feature sets. A discussion of the work follows in Chapter V, and Chapter VI concludes the thesis.

1.2 Background

1.2.1 Overweight and Obesity

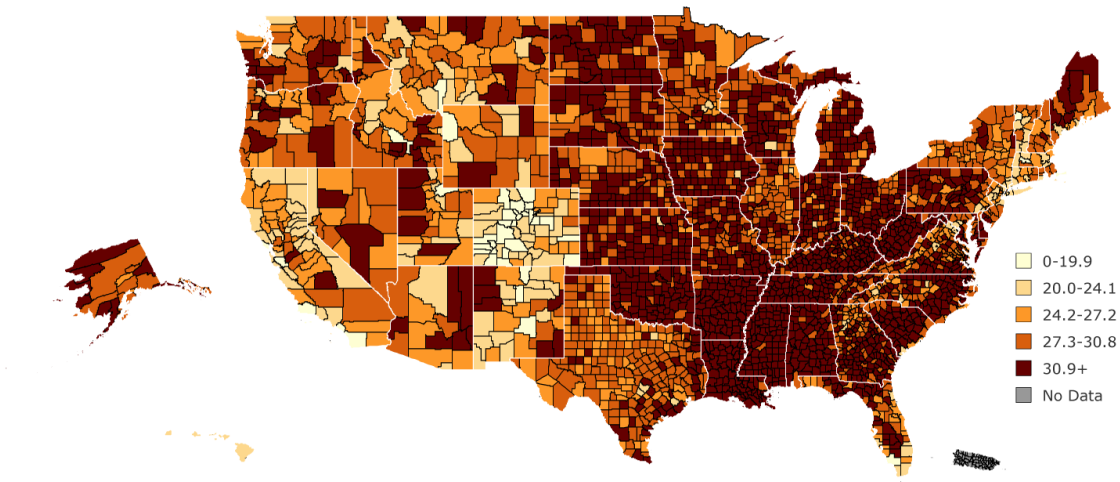

Figure 1.1: Percent of population classified as obese in the United States, by county [1]

Overweight and obesity are growing problems in the United States: over two thirds of the adult population are considered overweight, and 36.5% of adults aged 20 or older are considered obese [2]. An individual is classified as overweight if he or she has a Body Mass Index (BMI) - the ratio of mass in kilograms to the square of height

in meters - above 25 kg/m^2, and obese if his or her BMI is above 30 kg/m^2. A major risk factor for a number of other conditions and diseases, obesity has been linked to higher incidences of problems such as obstructive sleep apnea, type 2 diabetes mellitus, stroke, asthma, hypertension, hyperlipidemia, depression, and certain cancers [2–5]. These associated conditions combine to place an impressive burden on the healthcare industry, with estimates of the cost ranging from $147 billion to $209.7 billion annually (in 2008 dollars) [5, 6].

Many obese individuals attempt volitional weight loss through dieting, and those involved in controlled studies often see significant progress. Unfortunately, a series of neurobehavioral mechanisms are activated following a caloric restriction, some within 24 hours, that typically result in weight regain [7]. There are, however, a number of individuals that can maintain a notable decrease in weight: some estimates put this at 23% of dieters [8, 9]. To date, no factors have been definitively identified as predictive of future success.

1.2.2 Signal Processing

Signal processing involves taking as input some ordered collection or series of data and analyzing or modifying it to enhance a desired component or extract a particular characteristic. With applications in a broad and diverse set of scenarios, there are a wide array of available techniques. In the medical field in particular, signal processing has been used to analyze heart rate data, brain activity, physical motion, and numerous other measurements. The purposes and uses of these studies are just as varied, from peak detection and movement classification to motion artifact reduction and prediction of panic attacks [10–13].

One of the most-used methods in signal processing, the Fourier transform translates a signal from the time domain to the frequency domain [14]. By breaking down the input to a Fourier series, this method produces a set of peaks representing the

amplitudes of the original signal's constituent parts. By examining this decomposition, previously obscure information may more readily present. In addition, as the process is reversible (inverse Fourier transform), changes can be applied in the frequency domain to alter the signal in the time domain. For example, a low-pass filter only allows through those components of a signal's Fourier transform below a set frequency; this results in a time-domain signal with any high-frequency noise removed. This filtering, together with features extracted from the Fourier transform, has been used in countless medical studies. By applying various high- and low-pass filters, major components of noise can be removed from a recorded heart-rate signal [15]. The Fourier transform can also be used to detect abnormal breathing rates, epilepsy, and even certain cancers [16–18].

The wavelet transform, like the Fourier transform, involves the frequency domain. The main difference, though, is that while the Fourier transform looks at discrete functions set at specific frequencies (sines and cosines) to decompose the original signal, the wavelet transform uses a function with both time and frequency components (mother wavelet) [19]. While in a typical Fourier transform it is impossible to tell at what time a frequency is present, using a mother wavelet allows for the determination of which frequencies are present, and at what magnitude, for any given point in time. Also unlike the Fourier transform, which has a consistent scale across time and frequency, the wavelet transform has varying resolutions: the high frequencies are calculated at a superior temporal resolution than the low frequency components, which in turn have a better frequency resolution. While not as widespread, use of the wavelet transform has benefited numerous medical applications including signal noise reduction, heart rate variability characterization, gait analysis, and pacemaker design [20–23].

1.2.3 Machine Learning

Machine learning is a branch of computer science that explores algorithms meant to learn from data. Widely applicable, various implementations of this approach have been used to examine a wide range of problems. From protein-protein interactions and tumor classification to detecting oil spills in satellite images, machine learning algorithms have a large impact on current science [24–27].

In deep learning algorithms, multiple non-linear transformations are used to model the input data. These sets of non-linear processes are often regarded as 'black boxes' due to the fact that the user often does not see the workings of each individual layer [28]. As such, it is impossible to trace backwards any resulting classifications to determine why an instance was given a particular label. This in turn prevents any insight being gained into hidden patterns or underlying features that could have otherwise proven useful. In addition, for implementations with large sets of parameters, searching for ideal configurations is an extremely complicated and time-consuming process [29, 30]. An expert is often needed to help determine acceptable parameters, though this process is rarely deterministic and results in a conceptual barrier for non-experts [31]. In cases where finding the right set of parameters is crucial to the success of the algorithm, this could prove prohibitively restrictive.

Like deep learning, the random forest approach is also a popular machine learning method. While the result of the algorithm is a set of decision trees, the split rules in any one tree can often be too complex for any meaningful insights to be derived [32, 33]. Coupled with the number of trees typically generated by this approach, the reason behind the resulting classification for any input can be difficult to dissect. This again prevents the researcher from discovering or understanding the underlying patterns meant to be exposed by machine learning. The random forest method can also be extremely computationally intense, requiring the generation of a large number of potentially deep trees. For certain datasets, arbitrarily long trees are needed to

successfully split the training examples, a task that requires not only the computational power to generate the trees but also the space to keep them in memory [34]. In addition, while able to handle local noise, the random forest algorithm is highly susceptible to global noise [35]. For datasets suffering from this issue, such as those with nonzero baselines or wide background peaks, the random forest approach may produce useless classifications.

Support vector machines (SVMs) are another widely used machine learning approach. Unfortunately, as with the above methods, SVMs can be computationally prohibitive, both in terms of computational time and memory requirements [36, 37]. For datasets with large training groups, it may be impossible to store the required information and impractical to calculate it every time it is needed. The performance of SVMs is also subject to the underlying kernel used in the implementation, with some yielding strong interpolation results while others are better suited for extrapolation [38]. One of the major disadvantages to SVMs comes from the fact that they are inherently binary classifiers: for multiclass problems, extra steps must be taken to correctly classify an instance. This not only necessitates the construction of multiple binary classifiers, each subject to the computational disadvantages mentioned above, but it also raises the issue of how to combine the resulting outputs [39]. A true multiclass SVM could be used, but this would involve resolving an even larger optimization problem.

"Adaptive Boosting", or AdaBoost, is more a method of combining other machine learning algorithms than an algorithm itself. Like SVMs, AdaBoost was not designed to handle multiclass problems and generally grows more computationally intense with the introduction of more than two classes [40]. As the method relies on its constituent 'weak learners' having an error rate less than 50%, having multiple classes increases the likelihood of failure. For a dataset with complex classification rules, a weak learner with constrained computational resources - both time and memory - may

perform poorly, in turn breaking the general method [41]. In addition, AdaBoost is known to overfit the training data, resulting in poor performance when applied to testing or general instances [42, 43].

Computational time varies across the above methods, ranging from $O(N)$ to $O(N^3)$ [44]. This depends on both the algorithm and the kernel: Random Forest can go as $O(N^{1.5})$ or $O(MN\log(N))$, where M is the number of trees [44, 45]. These methods can be improved or optimized to achieve a computational complexity of $O(N)$, though some assumptions and reductions must be made to reach this level [46].

1.2.4 Data

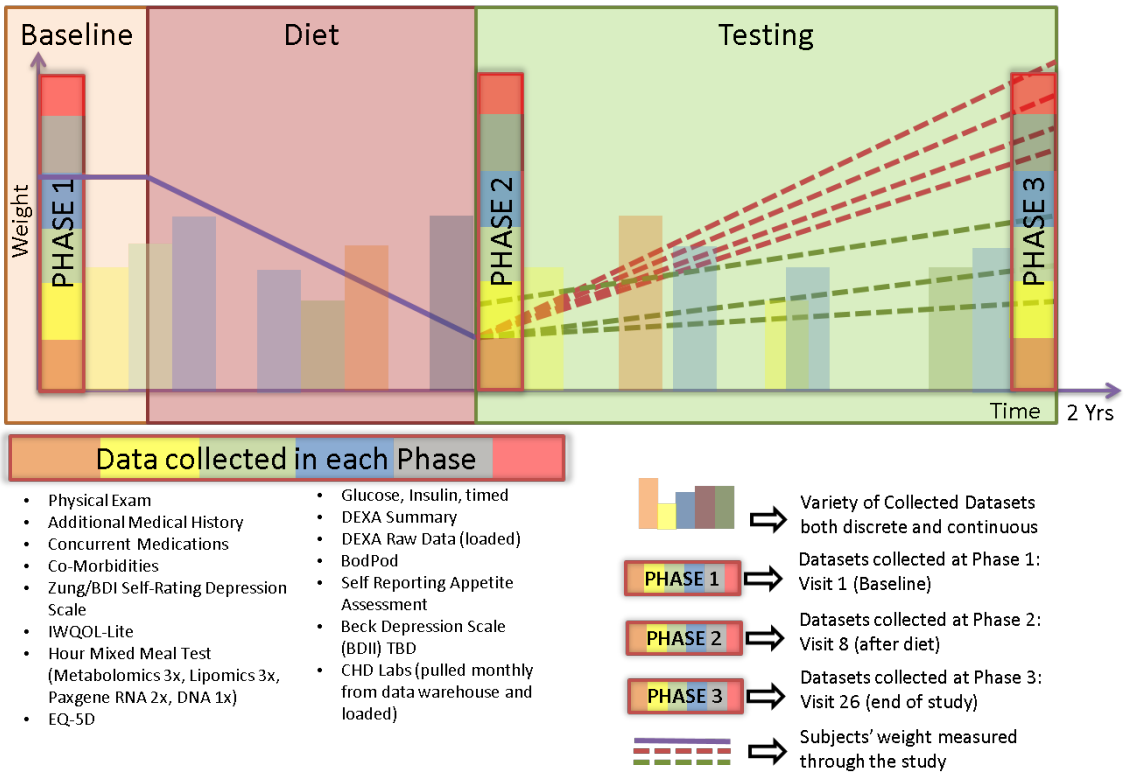

Figure 1.2: Data collected in and timeline of the full study

This research utilizes the databases created by a longitudinal study conducted by the University of Michigan Weight Management Program clinic (MWMP), a demon-

stration unit of the NIH-funded (grant DK089503) University of Michigan Nutrition Obesity Research Center. The MWMP is a two-year, multidisciplinary, multicomponent, structured lifestyle intervention which promotes 15% weight loss though intensive caloric restriction for the first 12 weeks. This is succeeded by interventions and routine follow-up aimed at supporting long-term behavioral change. All participants provided written informed consent and the study was approved by the University of Michigan Medical School's Institutional Review Boards (IRBMED, HUM0030088). As shown in 1.2, participants complete a wide array of clinical, psychological, and metabolic assessments at a baseline (Phase 1), after intensive weight loss over the course of 3-6 months (Phase 2), and at the end of two years (Phase 3). Included among these assessments are the measurements of an individual's movement profile, Resting Metabolic Rate (RMR), and fitness (peak rate of oxygen consumption, VO_2 peak). [1]

One complicating factor in analyzing this data lies in the timeline: because the tests are subject the participants' schedules, as well as those of the clinicians, the visits associated with phases 2 and 3 do not always occur on the ideal dates. Even within one phase, the different assessments may be spread over a significantly large time period. In addition, all aspects of the study are 'opt-in' which results in a sparse dataset. If an individual chooses not to complete a particular test, the corresponding section is left empty. This issue was largely handled at the clinical level, with leeway given in characterizing a particular visit as belonging to a certain phase. An additional difficulty presented in parsing the data. The results of each assessment were saved separately, spreading the measurements across a wide array of files and formats. As the tests were run by different clinicians over a number of years, the layout of any one file type also varied. Extensive scripts were developed to parse these disparate sources per participant and combine all relevant information into single, consistently

[1] This paragraph was previously submitted in Biwer et al. [47].

formatted files.

The data used in this research can be split into two categories: those collected at home and those collected in a laboratory or clinic setting. Details of the data from both categories will be presented in the chapters in which the algorithms used to analyze them appear. In addition to the tests used, a number of other assessments were performed. For example, each subject could complete a series of psychological assessments, undergo various blood tests and biopsies, or have a DEXA scan performed. While these provide a wealth of data, including them would have significantly reduced the already limited number of individuals who completed all relevant tests.

CHAPTER II

Signal Processing for Analyzing Activity Data [1]

2.1 Background

As medical monitoring devices continue to grow in complexity and shrink in size, both the number of possible concurrent measurements and the size of the observable population increase. These factors in turn result in a rise in the amount of data available for analysis, which is driving the need for new processing algorithms. The sheer volume of recorded values makes it difficult to process within a relevant timeline using conventional means, and the intricacies of some of the more obscure variables makes them difficult to interpret at all. Because of this, the need for novel signal processing algorithms that can detect and highlight underlying subtleties and features is becoming ever more apparent.

This problem, along with many others in the clinical setting where time-series measurements are to be analyzed, stands to benefit from advanced signal processing. A common practice in the medical field, various signal processing algorithms have been applied to a wide variety of situations. From heart rate monitoring to myoelectric signal classification, various techniques are used to analyze time series data. Feature extraction has proved to be an effective approach, commonly used in health applications, and involves calculating variables characteristic of the signal. Another

[1] Sections of this chapter were previously published in Biwer et al. [48].

approach, direct comparison methods, analyzes and compares raw signals simultaneously. The power, entropy, and average value of a signal are three common features used in signal processing, and the Pearson correlation coefficient is used when comparing signals and samples directly. The currently used methods, while effective in many applications, cannot always differentiate between two sets of similar signals in complex health applications, especially when the differences among signals are subtle. Indeed, the main shortcoming of existing models in predicting weight maintenance is a clear lack of effective computational approaches to mining available data. Instead, most of the existing predictive models of obesity are based on correlation of weight regain/loss with only a small amount of basic patient information, which result in models with limited predictive capabilities. For instance, a major factor that can help lead the personalization of treatments for obesity is estimation of the type and level of physical activities. However, the motion signals, when analyzed with the conventional signal processing methods, have failed to produce accurate and robust prediction results. The complexity of the data collected for any one patient, let alone that found in the collective data of a large study groups, demands more advanced computational techniques that can extract these subtle patterns and distinguish subclasses of weight loss and regain. This chapter describes a new approach in signal processing that can detect subtle changes in the behavior of complex signals, in particular motion time-series, and as such distinguish between patient cohorts with different clinical outcome. The proposed approach is based on an extended formulation of the persistent homology theory and introduction of a modified, semimetric version of the Hausdorff distance to analyze data in the feature space [49–55]. The proposed computational methods are applied, and their efficacy in differentiating between various levels of weight loss maintenance is demonstrated.

2.2 Materials and Methods

2.2.1 Data

As part of the study described in Chapter I, participants were routinely asked to wear a tri-axial accelerometer, hereafter referred to as an armband, that also measures galvanic skin response and near body ambient temperature (BodyMedia SenseWear armbands, http://www.bodymedia.com) for a period of 7 days, only removing the monitor to charge it while participating in water activities (e.g. showering, swimming). Each test yielded an individual file, which resulted in a large number of disparate data sources for each participant. Each individual, before testing began, gave his or her informed consent that the collected data be used for research purposes. In addition, after collection, all data files were de-identified and curated prior to analysis. This pre-processing also involved automatic and manual error-checking. Once all the relevant patient data were parsed, the files were spot-checked for inconsistencies. This was done by plotting various data values from numerous participant files and checking for outliers.

An advantage to this data lies in the fact that they can be passively recorded by the individual. By removing the complete reliance on self-reported activity levels, the armband allows for collection of signals while outside the clinic or laboratory without placing any undue responsibility on the wearer. Because of this, the data is not only cheaper to collect but their collection removes part of the burden previously placed on clinicians.

While worn, the armbands recorded the number of peaks in the accelerometer signal, once per minute, for each of three dimensions: transverse, forward, and longitudinal. These three numbers were summed each minute, resulting in a roughly week-long general movement profile for each individual. In this study, only participants who wore the device for at least 7000 minutes are included. This meant each

individual wore the device for at least 16.5 hours per day for the full week, or approximately 23 hours per day for five days. This number was chosen as it yielded as long a signal as possible while including as many participants as possible. Any participant for whom 7000 minutes of data were not recorded was excluded from this analysis, as well as those that had not yet progressed far enough into the study to have measurable results. Each included signal was cropped at the 7000-minute mark, resulting in a uniform length across all studied movement profiles. This was done for a number of reasons: the set length allowed for uniform mapping of the data without the need to stretch or skew the signals, preventing any interference from different time resolutions; no signal included more potential information than another, as they were all of equal length. A final inclusion requirement was that the participant be classified as a 'success' or 'failure'. As the specific aim of this study was to predict weight-loss maintenance, each participant was given a label based on weight loss and regain. If an individual failed to lose at least 15% of his or her starting body weight between Phase 1 and Phase 2, that person was labeled a 'failure'. Those who succeeded in achieving that goal, and who finished Phase 3 at a weight no greater than 90% of his or her starting weight, were labeled a 'success'. Any individual who completed the study without maintaining at least a 10% weight loss was considered a 'failure'. As a result, the individuals included in the study had all successfully completed at least Phases 1 and 2, and some had finished Phase 3. Accounting for all of the above criteria, 100 participants were included in this study. This cohort consisted of 36 males and 64 females with an average age of 50 ± 9 years old.

Next, the computational analysis based on persistent homology is described. A windowed formulation of persistent homology was used to extract characteristic features from the data from each participant. These features, represented as persistence diagrams, were then used to predict success. The ability of persistent homology-based features to predict success/failure was statistically analyzed, as described later.

2.2.2 Signal Analysis using Persistent Homology

Persistent homology is a broad mathematical theory, and one of its applications is examining how the characteristics of an object in a space change based on the spatial resolution used to examine the object. As the resolution changes, persistent homology features, representing the special characteristics of the object, quantify these changes. The transitions in these features can be studied to help develop a better understanding of the object. When applied to time series as objects, persistent homology can be used to extract features of a signal representing the changes in the characteristic patterns of variations observed in the signal at different resolutions [56–63]. Specifically, the persistent homology algorithm converts a signal into points scattered across a min-max plane. By treating each minimum in a time-series as the 'birth' of a feature and each maximum as a 'death' it is possible to examine the significance of a trend by the persistence of its corresponding min-max pairing. Larger differences between two extrema correspond to more pronounced variation, and any resulting points will be farther from the $y = x$ diagonal. Conversely, a point closer to the diagonal represents a smaller magnitude of change and is more likely to be noise. The resulting min-max plot, or 'persistence diagram', represents the characteristics of the input sequence and can be used to compare the differences between the patterns and variations of signals. This information can also be visualized in a 'barcode' format, as described in Ghrist [64]. The process, including the derivation of the persistence diagram, is as follows:

Suppose that f is a real-valued function on the discrete set $\{1, 2, \ldots, n\}$. To make notation convenient, define $f(0) = \infty$ and $f(n + 1) = -\infty$. Also, modify f to define a function $\widetilde{f}$ defined by $\widetilde{f}(i) = f(i) + \varepsilon i$ where $\varepsilon > 0$ is infinitesimal. We define the function-value ordering $\sqsubseteq$ on $\{0, 1, \ldots, n + 1\}$ as follows. If $0 \leqslant i, j \leqslant n + 1$ then define $i \sqsubseteq j$ if $\widetilde{f}(i) \leqslant \widetilde{f}(j)$. Equivalently, $i \sqsubseteq j$ if $f(i) < f(j)$, or $f(i) = f(j)$ and $i < j$. The relation $\sqsubseteq$ on $\{0, 1, , \ldots, n + 1\}$ is a total ordering. So for all i and j:

1. $i \sqsubseteq i$;

2. $i \sqsubseteq j$ or $j \sqsubseteq i$;

3. if $i \sqsubseteq j$ and $j \sqsubseteq i$, then $i = j$;

4. if $i \sqsubseteq j$ and $j \sqsubseteq k$ then $i \sqsubseteq k$.

We also write $i \sqsubset j$ if $i \sqsubseteq j$ and $i \neq j$. Define the set of local minima and maxima by

$$E_{\min} = \{i \mid 1 \leqslant i \leqslant n,\ i \sqsubset i-1, i \sqsubset i+1\} \tag{2.1a}$$

$$E_{\max} = \{i \mid 1 \leqslant i \leqslant n,\ i-1 \sqsubset i,\ i+1 \sqsubset i\} \tag{2.1b}$$

respectively.

Lemma II.1. *The sets $E_{\min}$ and $E_{\max}$ have the same number of elements.*

Proof. It is not hard so see that the smallest element of the set of extrema $E = E_{\min} \cup E_{\max}$ with respect to the ordering $\leqslant$ lies in $E_{\min}$. Also, the largest element lies in $E_{\max}$. It is also elementary to see that local maxima and local minima alternate. So the number of local maxima and local minima is the same. □

By the lemma, there exists $r, a_1, a_2, \ldots, a_r, b_1, b_2, \ldots, b_r$ such that $E_{\min} = \{a_1, a_2, \ldots, a_r\}$ and $E_{\max} = \{b_1, b_2, \ldots, b_r\}$ such that $a_1 \sqsubset a_2 \sqsubset \cdots \sqsubset a_r$ and $b_1 \sqsubset b_2 \sqsubset \cdots \sqsubset b_r$. The proof of the lemma shows that there are permutations τ and γ in the symmetric group S_r such that

$$a_{\tau(1)} < b_{\gamma(1)} < a_{\tau(2)} < b_{\gamma(2)} < \cdots < a_{\tau(r)} < b_{\gamma(r)}.$$

Lemma II.2. *We have $a_i \sqsubset b_i$ for all i.*

Proof. It is clear that $a_{\tau(i)} \sqsubset b_{\gamma(i)}$ for all i. For $j \leqslant i$, $a_{\gamma^{-1}\tau(j)} \sqsubset b_j \sqsubseteq b_i$. So at least i of the a's are smaller than b_i with respect to $\sqsubset$. This implies that $a_i \sqsubset b_i$. □

If $a, b \in \mathbb{R}$ then define

$$(a,b) = \begin{cases} \{x \in \mathbb{R} \mid a < x < b\}, & \text{if } a \leqslant b \\ \{x \in \mathbb{R} \mid b < x < a\}, & \text{if } a > b \end{cases} \tag{2.2}$$

We define a permutation $\sigma \in S_r$ and sets $U_1, U_2, \ldots, U_r$ inductively as follows:

$$U_k = \left\{ i \,\middle|\, 1 \leqslant i \leqslant r,\ a_i \sqsubset b_k,\ i \notin \{\sigma(1), \sigma(2), \ldots, \sigma(k-1)\} \right\} \tag{2.3}$$

and

$$\sigma(k) = \max\{i \mid i \in U_k \text{ and for all } j \in U_k \text{ with } j < i,\ a_j \notin (a_i, b_k)\}. \tag{2.4}$$

By Lemma II.2, the set U_k is nonempty.

Definition II.3. The persistence diagram associated to the function f is

$$\{(f(a_{\sigma(1)}), f(b_1)), (f(a_{\sigma(2)}), f(b_2)), \ldots, (f(a_{\sigma(r)}), f(b_r))\}.$$

The mathematical complexity of the persistent homology approach formulated above might mask its great conceptual potentials to analyze complex signals. Below is a graphical illustration and explanation of this method. As shown in Figure 2.1, first a given signal is plotted and all local extrema (maxima and minima) are identified. A square plot in the min-max plane is generated and placed such that the vertical axes align between the two graphs, and the line $y = x$ is drawn (Figure 2.1a). Starting with the smallest minimum (in this case, point 1), a mark is placed on the min axis of the second graph corresponding to the minimum's y value (Figure 2.1b). Each local extremum is considered, moving from smallest to largest, and for each minimum another mark is placed (Figure 2.1c). When a maximum is encountered, that point's y value is paired with the mark from the most recent unpaired minimum, as long as there is no other maximum between the two points. If such a maximum exists,

the second most recent minimum is considered, and the process continues until a suitable pairing is found (Figure 2.1d). The next extremum is then considered, and the algorithm proceeds until all points are paired (Figure 2.1e). In the event that there are more minima than maxima (or vice versa), the point with the largest x value in the larger set is dropped. Once completed, the points in the min-max plane constitute the signal's persistence diagram (Figure 2.1f). As it can be seen, starting from a time-series, persistent homology creates a pattern of dots in the max-min plot that represent the variations of the signal.

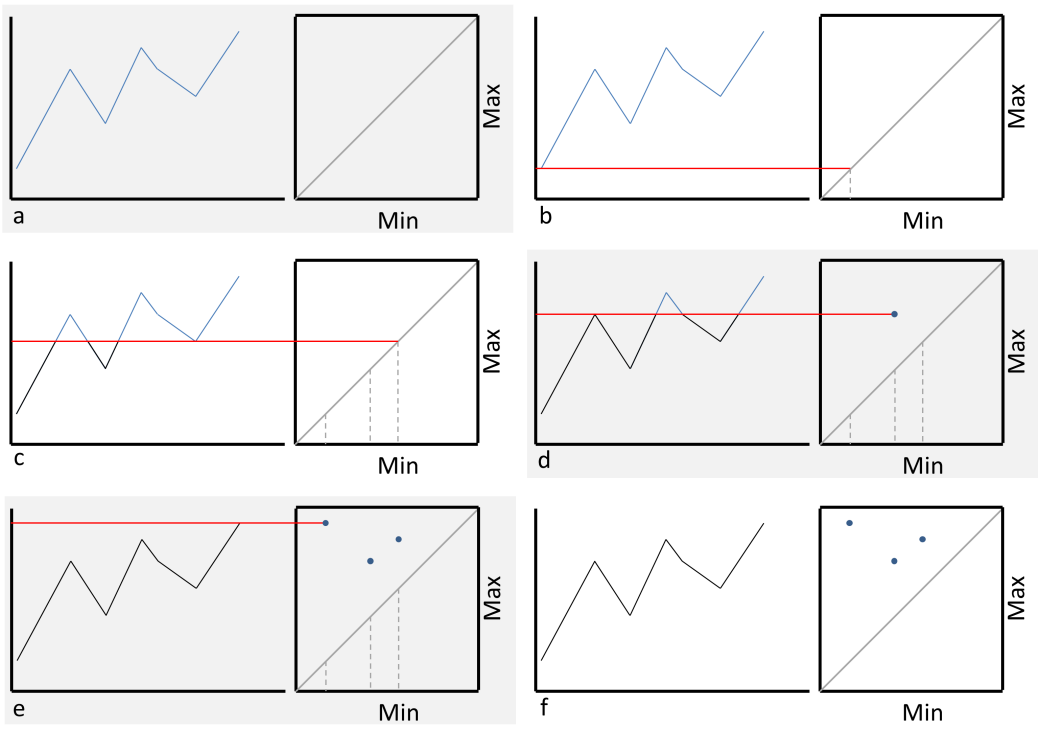

Figure 2.1: A visual representation of persistent homology

These min-max patterns represent how the variations in signals can be used, as described later, to effectively distinguish between different types of signals (e.g. signals representing 'success' and those representing 'failure'). Next is a description of the method for distinguishing between different patterns/signals by using a measure that quantifies the disparity between different min-max plots.

2.2.3 Assessment of Feature Space using Modified Hausdorff Semimetric and Wasserstein Distance

The persistent homology method can be used either as a stand-alone analysis tool or as a comparative metric. For the former case, the persistence diagram can be examined and features extracted, and for the latter two persistence diagrams can be compared in a number of different ways. For this study, persistent homology was used to compare different armband activity signals. However, calculating and comparing the persistence diagrams for the full 7000 points from each input is not only computationally expensive but also masks some important patterns in parts of the signals due to the averaging effect. Moreover, for such long signals, not only would the persistence diagrams be extremely dense but any order to the signals would be lost: points generated by pairings of extrema from the beginning of one signal could be near points generated by extrema at the end of the second signal. In addition, the dots themselves may result in inaccurate interpretation or analysis of the signal. For instance, pairing a minimum from near the beginning of the signal with a maximum from near the end would be relating two otherwise independent measurements: the activity from the first point would be entirely separate from the activity that generated the second. As such, comparing the persistence diagrams of two long signals would yield little valuable information. To solve this problem the armband signals were first broken into windows and the algorithm was applied to each window. This had the added benefit of decreasing the computational time of the implementation by a factor of over 300.

This study involves the design and implementation of a windowed based approach to persistent homology to address the above mentioned issues. Specifically, each 7000-point signal was broken into 350 windows, each containing 20 points. This was done by simply splitting the original signal into equally-sized standalone segments using a rectangular window: no overlap or tapering was used. The window size was chosen

because it allowed for reasonable variation within a window while at the same time ensuring that any two paired points would be closely related in time. A persistence diagram was calculated for each window, and corresponding windows from two signals were compared (i.e. the first window from each signal was paired, followed by the second, etc.). An overview of the process can be seen in Figure 2.2.

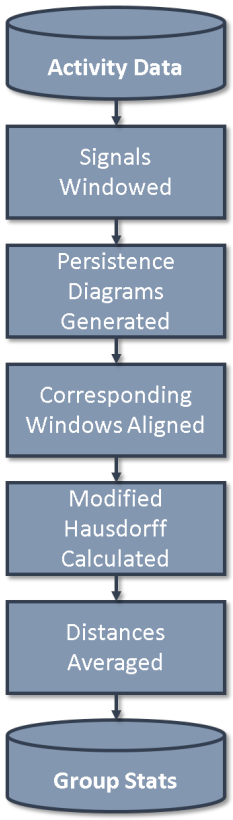

Figure 2.2: Overview of the windowed persistent homology method

Various metrics exist for calculating the distance between two sets of points, including the Hausdorff distance and the q-th Wasserstein distance. The latter is defined in Kerber et al. [65] between two sets A and B as:

$$W_q(A,B) = \left[\inf_{f:A\to B} \sum_{a\in A} ||a - f(a)||_\infty^q \right]^{1/q} \tag{2.5}$$

where f is the set of all bijections $A \to B$. When $q = 1$, this metric reduces to the minimum possible sum of the distances between each point in A and its corresponding

point in B. By allowing points to map to the $y = x$ line, any contribution from noise is minimal. The Hausdorff distance d_H is calculated as:

$$d_H(A, B) = \max\{\sup_{a \in A} \min_{b \in B} d(a, b), \sup_{b \in B} \min_{a \in A} d(a, b)\} \tag{2.6}$$

where $d(a, b)$ is the Euclidean distance between points a and b. However, this method is quite sensitive to outliers: one anomalous point in either set could greatly skew the measured value. Due to the stochastic and noisy nature of the data being examined, the metric used to compare two persistence diagrams had to be tolerant of such deviations and outliers; if it was not, one large peak caused by on outlier could alter the distance. This led us to the use of a modified, semimetric version of the Hausdorff distance [55]. Replacing the inner suprema with an average gives:

$$d_{mH}(A, B) = \max\left\{\frac{1}{|A|}\sum_{a \in A} \min_{b \in B} d(a, b), \frac{1}{|B|}\sum_{b \in B} \min_{a \in A} d(a, b)\right\} \tag{2.7}$$

This metric is much more tolerant of outliers as they are included as one part in a general sum and not the only representative number. However, as noted in Dubuisson and Jain [55], this version is not a true distance metric as it does not satisfy the triangle inequality. As such, because it satisfies the other distance metric requirements, this comparison qualifies as a semimetric.

2.3 Results

To begin, the power of each raw armband data was calculated (Table 2.1). This rather simple and intuitive feature is often used in analysis of activity signals as a main characteristic number describing the data. In this study, however, there is no measurable difference between the power of a participant labeled as a failure and that of a participant labeled as a success ($p = 0.326$). The sample size was 100, of which

79 were labeled 'failures' and the remaining 21 were considered 'successes'.

Table 2.1: Standard signal features extracted from activity data

	Label	**Average**	**Standard Deviation**
Power ($p = 0.326$)	Failures	196205	56942
	Successes	182888	46034
Entropy ($p = 0.608$)	Failures	0.6560	0.1433
	Successes	0.6732	0.1037
Average ($p = 0.262$)	Failures	289.9	57.1
	Successes	274.3	53.0

The entropy and average of signals, also popular in assessment of activity signals, were calculated and analyzed as well (Table 2.1). Entropy as a measure of disorder and information can often distinguish between functional classes of data, in particular when dealing with biomedical signals, and is heavily used in the signal processing literature. However, as in the case of power, there was no significant difference between the failure and success groups ($p = 0.608$ and $p = 0.262$, respectively).

As a final check using current established methods, correlation coefficient calculation was used to analyze the data. For this, each signal was compared to every other signal in the pool and a correlation value was obtained. When armband data from two individuals labeled as failures were compared, the resulting signal was placed into a 'failure vs failure' group (N = 3081); likewise, when the movement profiles of two successes were compared, the result was placed into a 'success vs success' group (N = 210). When comparing two 'failure' signals, the average correlation coefficient was 0.0712, while the average for comparing two 'success' signals was 0.0992. While the standard deviations were relatively high, as shown in Table 2.2, there was a statistically relevant difference between the groups ($p = 0.006$).

Comparing the 'failure vs success' group (N = 1659) to the 'failure vs failure' group yields another statistically significant difference ($p = 0.0004$), but comparing it to the 'success vs success' group does not ($p = 0.1968$). This indicates that the failures share less intra-group similarities than do the successes, but the second two

Table 2.2: Pairwise correlation analysis results

Correlation ($p = 0.006$)	**Average**	**Standard Deviation**
Failure vs Failure	0.0712	0.1431
Failure vs Success	0.0864	0.1376
Success vs Success	0.0992	0.1158

cases are indistinguishable under this metric.

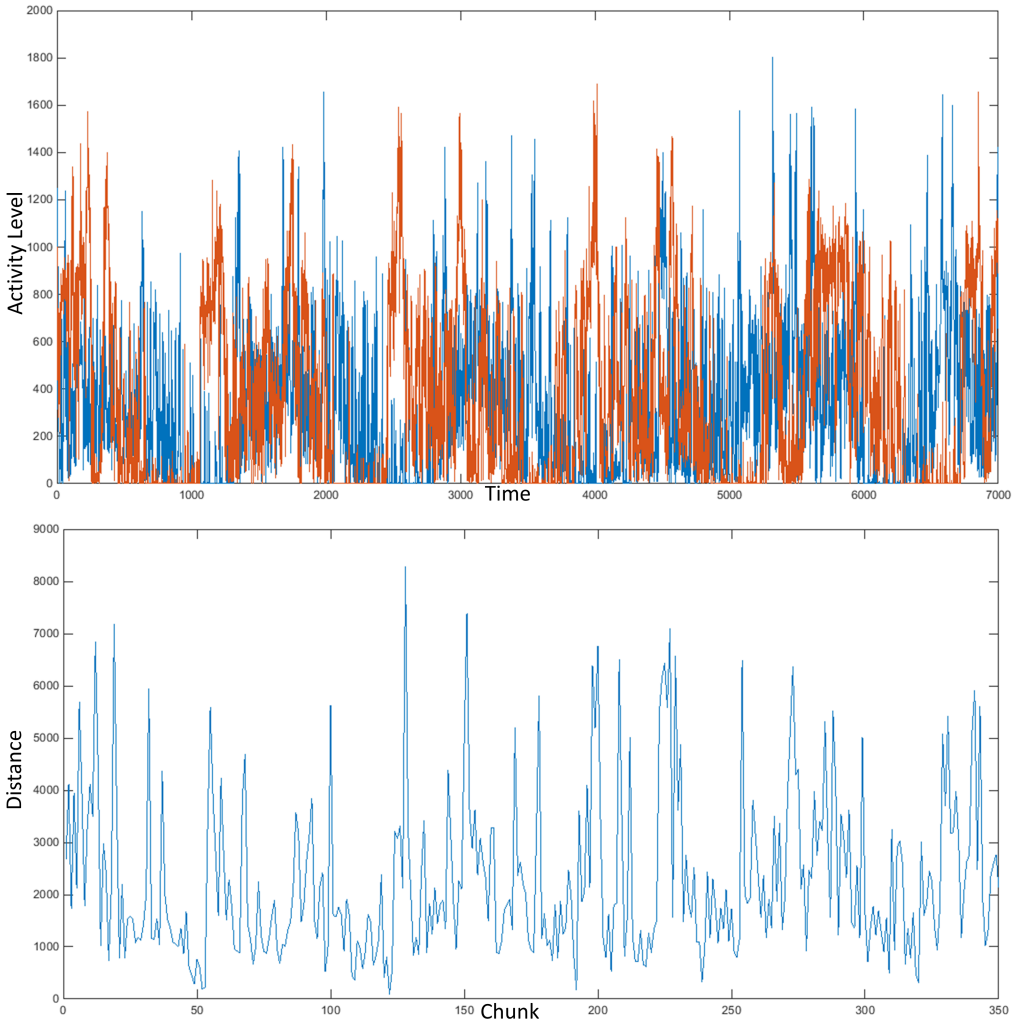

Figure 2.3: Activity signal comparison using windowed persistent homology and a modified Hausdorff semimetric

Next, the armband data of each of the 100 participants was compared to every other file using the proposed windowed persistent homology method. Persistence diagrams were generated for each of the 350 windows extracted from a signal. These

plots were compared to corresponding plots from a second signal using the modified Hausdorff semimetric, and a single number was noted for each window. The result of the algorithm, when applied to two armband/activity files, was a new signal with a length of 350 points, corresponding to the distance between the two input plots over time. The top of Figure 2.3 shows two armband signals plotted together, one drawn in red and the other in blue. Plotting the measured distance between two corresponding windows over the course of the analysis yields the signal shown in the bottom of Figure 2.3.

The average value of the resulting signal was placed into a group based on the input signals, again separating 'failure vs failure', 'failure vs success', and 'success vs success'. The average of these averages was then calculated and this was used as the characteristic value for each group (Table 2.3). Using an unpaired t test, the analysis showed that there is a statistically significant difference between not only the two main groups ($p = 1.505 \times 10^{-23}$) but between any two of the three ($p = 1.661 \times 10^{-28}$ and $p = 5.715 \times 10^{-9}$ for 'failure vs failure' vs 'failure vs success' and 'failure vs success' vs 'success vs success', respectively). It should be noted that the smaller average distance between successes when compared to that between failures is further strengthened by the correlation analysis: while the coefficients were small, the successes tended to be more highly correlated with one another than did the failures.

Table 2.3: Results of applying windowed persistent homology with a modified Hausdorff semimetric to activity data

Per. Hom. ($p = 1.505 \times 10^{-23}$)	**Average**	**Standard Deviation**
Failure vs Failure	378.27	64.79
Failure vs Success	356.99	58.59
Success vs Success	332.58	41.29

These results were consistent across variations in window size: values from 15 to 25 were also tried, with p-values no worse than an order of magnitude higher ($p = 1.170 \times 10^{-8}$ for 'failure vs success' vs 'success vs success' with a window size

of 16); some comparisons were more significant. Additionally, the method was implemented on 'blind' files in which the labels were randomly generated. With this random distribution there were no statistically significant differences between the groups (average $p = 0.113$ over 100 trials for the main pairing), further reinforcing the notion that the presented method is capturing some difference in the underlying structures of the signals.

Finally, the same approach was applied, but this time the q-th Wasserstein true distance metric was used to compare the persistence diagrams (q equal to one; code provided by Kerber et al. [65] was used). As shown in Table 2.4, the results are again statistically significant when comparing the 'failure vs failure' group to the 'success vs success' group ($p = 1.241 \times 10^{-5}$). While comparing the 'failure vs failure' cohort to the 'failure vs success' set also yields a significant difference ($p = 6.181 \times 10^{-7}$), it should be noted that both p-values are larger than their counterparts obtained using the semimetric. In addition, this true metric does not detect a measurable variation between the 'failure vs success' and 'success vs success' groups ($p = 0.012$). When run with randomly generated labels, the results are once again not significant for any combination of groups (average $p = 0.084$ over 100 trials for the main pairing).

Table 2.4: Results of applying windowed persistent homology with a q-Wasserstein distance to activity data

q-Wass. ($p = 1.241 \times 10^{-5}$)	**Average**	**Standard Deviation**
Failure vs Failure	711.88	63.17
Failure vs Success	702.63	56.36
Success vs Success	692.48	44.47

Despite the lack of complete statistical significance between all pairs, the pattern of higher intra-group similarities in the successes than the failures is continued. This, combined with the results of the 'blind' files, lends even more credence to the claim that there is an underlying disparity between the behaviors of the two groups.

2.4 Discussion

Figure 2.4 shows an example comparison from each of the three groupings. As shown in the plots, the general movement profile recorded by the armband sensors varies more heavily within the failure group than in the success group. Intuitively, this indicates that those destined to fail behave in a variety of different ways while those who will lose weight and maintain the loss share a more unified pattern of behavior. The general topographical structures present in the signals represent these overall patterns of activity and are captured in the persistence diagrams; in this sense, persistent homology is ideally suited to uncovering the underlying differences. Because of the larger variation present across the movement profiles of the 'failures', the signal-to-signal comparisons yield consistently higher values of both the average distance and the corresponding standard deviation across all metrics (true and semi-) used above.

In utilizing the presented persistent homology algorithm, a set of connected components from the time-series data are extracted [66]. In looking at this homology group, the underlying patterns of each individual's short-term behavior are exposed. Intuitively, this shows that the types, frequencies, and amplitudes of movement vary between each group, not just in general trends but also in minute-to-minute fluctuations. Future studies of this phenomenon could lead to discoveries pertaining to physical movement and how it contributes to and informs future weight loss success.

While further analyses with more participants would greatly help strengthen and validate these results, the initial implications are twofold. First, the clear disparity between the two groups indicates that there is a measurable difference in the movement profiles of those that will lose weight and keep it off versus those that will not lose any or, after losing weight, regain a substantial amount. By measuring this contrast a patient could potentially be classified as a 'success' or a 'failure' before even beginning a diet, in turn leading to more effective and individually tailored interven-

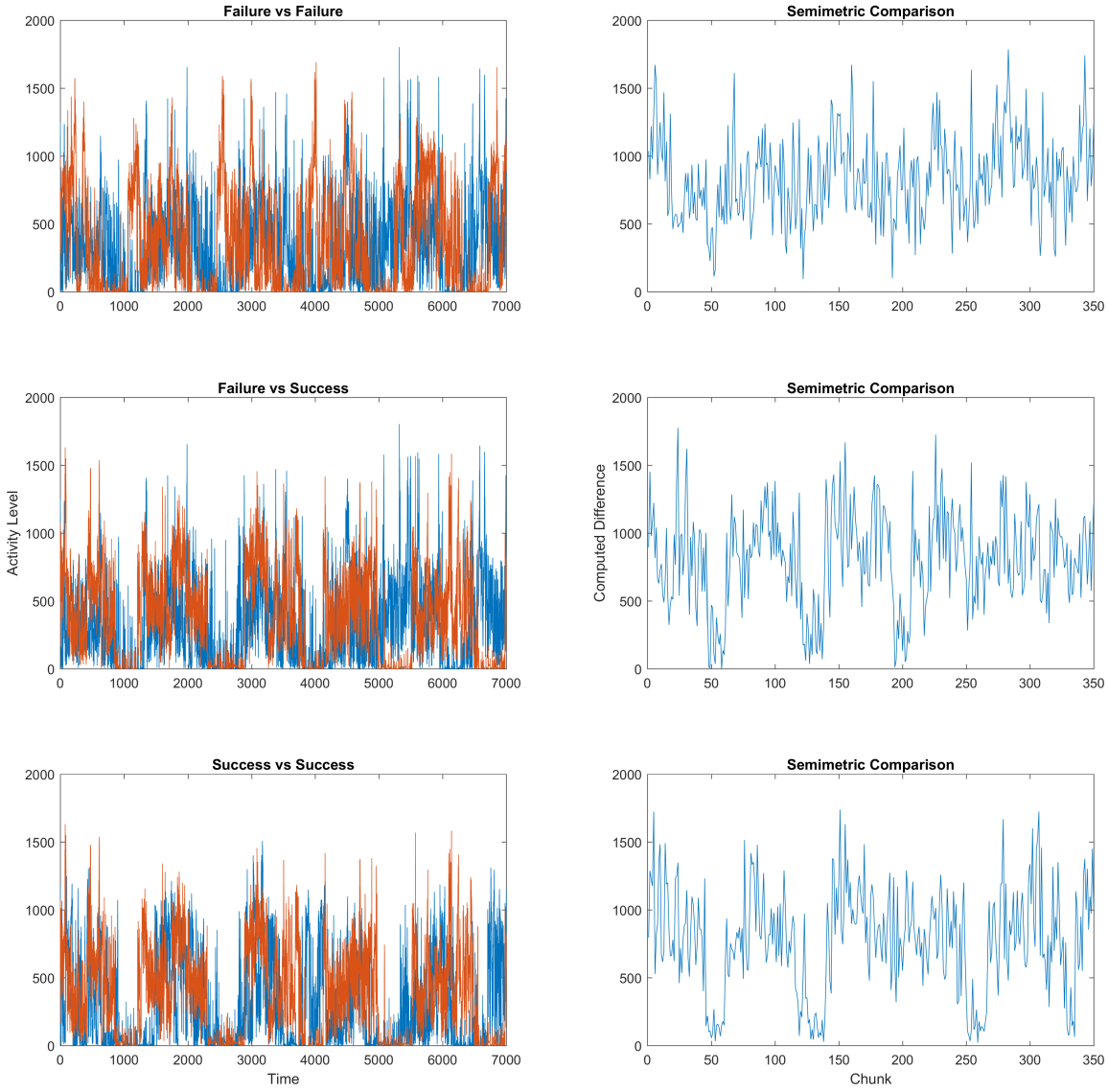

Figure 2.4: Example signal comparisons between different classes

tions. This would help to greatly ease the economic costs associated with overweight and obesity, as well as their related diseases. It would also save both the clinician and the patient time spent pursuing a course of action likely to produce unsatisfactory results, instead allowing them the option to first pursue alternatives. Secondly, and more immediately evident, the results indicate that the windowed persistent homology method, coupled with the modified Hausdorff semimetric, is capable of detecting subtle, underlying differences between signals. This method could potentially be used in other clinical settings where a deeper analysis of a complex signal would result in improved care, as well as other signal processing applications. Relatively tolerant of noise and sampling frequency, the presented algorithm can be easily applied to short and long time-series alike, drawing out features from the signal useful in exposing subtle differences.

2.4.1 Future Work

While the exact physical characteristics measured by the persistence diagram remain unclear and will be closely examined in future work, it can be said that the patterns representing more scattering in the persistence diagram represent higher levels of physical activity. In addition to further exploring the physical and physiological implications of an individual persistence diagram, there are a number of modifications to the implemented algorithm that will be investigated in future work. For instance, in this analysis, the inputs were blindly compared. In future work, a set of alignment procedures in the pre-processing steps will be implemented. Syncing time of day or sleep/wake cycles between two armband signals before the persistent homology algorithm is applied could help reduce any noise associated with comparing across states (e.g. one participants sleep data with another's morning routine). Another route for future investigations will be to implement a dynamic windowing method: by altering the length of each window based on the number of included

extrema, the resolution of the algorithm in areas of high activity (e.g. exercise) can be improved without sacrificing accuracy in low-movement periods (e.g. sleep). A third future improvement to the analysis involves error-checking the edge cases. If a window boundary splits a monotonically increasing or decreasing section of the signal, a false maximum and minimum are formed on either side. By implementing a check on the next point outside any given window, the number of artificial extrema can be reduced, thus minimizing inaccurate pairings and noise in the persistence diagrams.

2.5 Summary

To predict *a priori* weight loss/maintenance success in overweight or obese individuals by applying information learned from one week of simple, noninvasive measures would heavily impact the healthcare industry. With such a high prevalence rate in the country, both the economic burden and the time spent treating overweight, obesity, and their related diseases could be drastically reduced. By facilitating more tailored and individualized treatment, which may include modifying an existing program, identifying alternative modalities, or focusing on other patient issues, countless hours could be saved for both the patient population and the clinicians. The results presented in this chapter indicate, through the use of a novel computational method, a measurable contrast between the group of participants able to maintain weight loss and the group unable to do so. By using the windowed persistent homology method defined above and the modified Hausdorff semimetric, a physician could determine whether or not a specific intervention would be effective for a given patient. This project demonstrates the effectiveness of the novel signal processing method and the potential impact it can have on clinical decision making and patient care.

CHAPTER III

Machine Learning Applied to Physiological Data to Predict Future Prescription Medication Use [1]

3.1 Background

As mentioned in Chapter I, two major in-clinic tests often used in physiological assessments are the Resting Metabolic Rate (RMR) and VO_2 peak exams. These provide valuable information on an individual's basal metabolic rate and level of fitness, but require a trained clinician to administer. While this provides for more accurate results, it also places a temporal burden on the clinician. Moreover, quantitative interpretation of the resulting data, considering the complexity of the time series produced by these machines, poses a challenge. As such, applying signal processing techniques to analyze this data can provide invaluable clues to clinicians that would have otherwise gone unnoticed.

By applying current methodologies in signal processing and machine learning, new insights can be gained in studying this problem. Already commonplace in the medical field, numerous situations have benefited from the use of signal processing algorithms. For example, myoelectric signal classification and heart rate monitoring are two areas in which modern techniques have been successfully used [67, 68]. In particular, feature

[1] Sections of this chapter were previously submitted as Biwer et al. [47].

extraction has proven quite effective: in this process, an input signal is analyzed and various characteristic measurements that describe it are obtained. This allows for an extremely long, dense time-series to be condensed into a significantly smaller dataset that can then be fed into various machine learning algorithms. By further analyzing the extracted features to determine which possess meaningful information - that is, which features are better suited to help predict the labeled outcome - the feature set can be reduced to an even more compact size. This can not only increase the accuracy of the machine learning techniques by removing unwanted noise, but it can also greatly decrease the computational time required to build the models. When applied to physiological data, this overall process can yield important and clinically relevant predictions based on simple inputs. This chapter presents a computational method to analyze clinical data, namely physiological measurements from two different study phases, and predict the long-term impact of diet-induced weight loss on reduction or increase in number of medications.

3.2 Methods

3.2.1 Data

Both RMR and VO_2 peak tests are completed in the morning following an overnight fast ($\geqslant$ 10 hours), with the RMR assessment completed by 10am to minimize the effect of circadian rhythm [69, 70]. Height and weight are measured in light clothing and stocking feet. Baseline height is utilized for all subsequent assessments. Participants are escorted in to a darkened room and rested on a bed for one hour. The bed is flat but includes an adjustable 30-degree incline for the participant's chest and head. Individuals are provided one pillow and covered with two light blankets. Participants are provided additional pillows if desired and are permitted to remove blankets at will. They are instructed to remain still, quiet, to relax but not to sleep,

and are closely monitored throughout the entire test. If sleep is suspected, study personnel first create noise - i.e. kick two office chairs together, loudly tap a pen against a countertop - and if participant does not stir lightly tap the individual's feet. At thirty minutes, a canopy is placed over the participant's head and upper chest, permitting measurements of O_2 consumption and CO_2 exhalation for the ensuing half hour. Per manufacturer protocol, all fan and canopy adjustments are completed within ten minutes of the initial canopy placement. Upon completion of the RMR assessment, participants are encouraged to rise, use the restroom, and change into exercise attire in preparation for the VO_2 assessment. Shades are opened, lights turned on, and the metabolic cart is re-calibrated to suit exercise assessment. Participants complete a modified Bruce treadmill protocol. Those with joint or other concerns that limit the ability to ambulate on an incline instead complete a modified Balke treadmill protocol. An easy, 'dummy' stage is included in the beginning of each protocol to enable successful completion by even the least fit of study participants. Both protocols feature ramped stages which increase in incline, speed, or both. The test is stopped at volitional exhaustion or if study personnel deems it necessary to halt for safety reasons. All tests are conducted by CPR/AED certified study personnel.

A number of signals (Table 3.1) are recorded continuously during both RMR and VO_2 peak assessments. Per manufacturer recommendations, smoothed averages are created every 60 seconds during the RMR assessment and every 30 seconds during the VO_2 assessment, resulting in low-frequency time-series for both. The results of each procedure are saved to standalone files, one per test.

Figure 3.1 shows a schematic of the methodology used in this study. After collection, all data files were deidentified. They were then aggregated and parsed by individual, with each participant's data saved into one distinct file. This allowed for easier analysis, as all recorded tests were in one place. It also served as a means to check for errors: when the parser encountered a problem, it was examined by hand.

Table 3.1: Description of signals captured during RMR and VO_2 tests

Signal	**Tests**	**Description**
F_ECO_2	RMR, VO_2	CO_2 in expired air (%)
F_EO_2	RMR, VO_2	O_2 in expired air (%)
METS	RMR, VO_2	Energy cost of physical activity
REE	RMR	Energy expended at rest
RQ / RER	RMR, VO_2	CO_2 produced / O_2 consumed
RR	VO_2	Respiratory Rate (breaths / minute)
Time	RMR, VO_2	Seconds since test start
Treadmill Elevation	VO_2	Treadmill incline (% gradient)
Treadmill Speed	VO_2	Treadmill pace (miles per hour)
V_E	RMR, VO_2	Minute Volume (L/min)
VCO_2	RMR, VO_2	CO_2 processed (L/min)
VO_2	RMR, VO_2	Oxygen processed (L/min)
VO_2/KG	RMR, VO_2	Normalized oxygen processed (ml / (kg $\times$ min))

Once this process was complete, the files were also checked for additional errors by plotting the same values across multiple participants and looking for outliers. For the purposes of this study, only the datasets from phases 1 and 2 were used for each of the RMR and VO_2 tests; inclusion of phase 3 data would negate the benefit of early prediction.

Both treadmill elevation and speed were included in the processed signals for the VO_2 peak assessment as they contained information on the protocol used in the test (e.g. the rate at which the speed / incline increase). Respiratory rate was also recorded during this test but not the RMR assessment. Similarly, resting energy expenditure was included in the RMR output and not the VO_2 peak output. As the exercise test continued until exhaustion, the length of the recorded signals differed for each participant. To compensate for this variance, each signal was interpolated using a cubic spline from which thirty equidistant points were pulled. This resulted in signals of identical length for both tests (RMR and VO_2 peak; the former was a

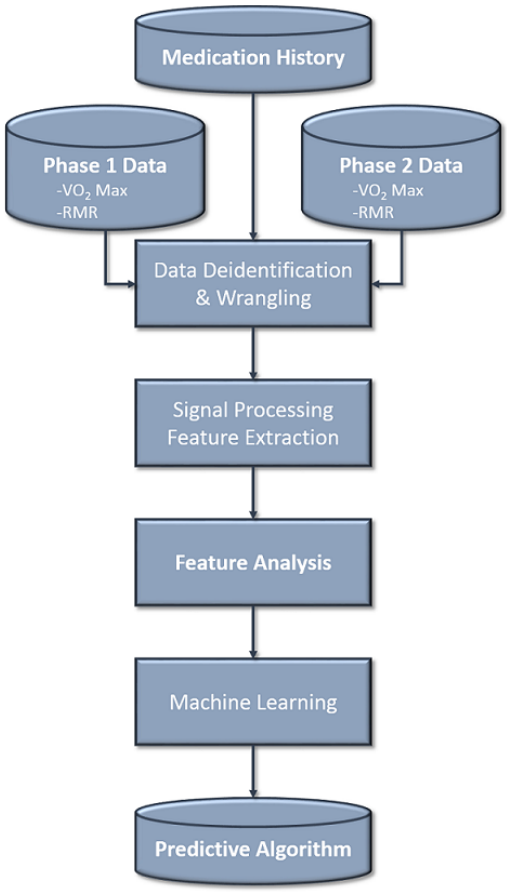

Figure 3.1: Data analysis pipeline

static thirty points regardless of participant) across all subjects. When completed, a wide variety of features were extracted from each short signal from each individual (Table 3.2). These were chosen to capture the general trend and characteristics of each set of points.

3.2.2 Feature Calculation and Normalization

The extracted features can be grouped into two categories: statistical and algebraic. For the former, the maximum and minimum values were noted for each signal, along with the average value, the range of the data, and the 25^{th}, 50^{th}, and 75^{th} percentiles. These allowed for a small but detailed description of the data. The algebraic features extracted were more intricate. The first value calculated was the total

Table 3.2: Descriptions of features extracted from RMR and VO_2 signals

Feature	**Description**
Max	The global maximum recorded during the test
Min	The global minimum recorded
Average	The average value of the signal (mean)
Range	The difference between the maximum and minimum values
25^{th}	The 25^{th} percentile value
50^{th}	The 50^{th} percentile value (median)
75^{th}	The 75^{th} percentile value
Number of turns (C_e)	The number of local extrema
Fit	The slope of the best-fit line
Fit residuals	The residuals of the best-fit (R^2)
First-half fit	The slope of the best-fit line for the first half of the data
First-half fit residuals	The residuals of the first-half best-fit line
Second-half fit	The slope of the best-fit line for the second half of the data
Second-half fit residuals	The residuals of the second-half best-fit line

number of local extrema, or number of times the series changed direction (C_e). For a given signal, first remove any consecutive repeated values: by consolidating any horizontal sections to a single point, it becomes easier to detect local extrema that have been stretched into plateaus. With this reduced sequence composed of points $s_1, s_2, ..., s_N$, define the second order geometric differential d at point i as:

$$d_i = (p_{i+1} - p_i)(p_i - p_{i-1}) \tag{3.1}$$

Using this, calculate the number of local extrema C_e as:

$$C_e = -\frac{1}{2}\sum_{i=2}^{n-1}\frac{d_i}{|d_i|} + \frac{n-2}{2} \tag{3.2}$$

The slope of the line of best fit for a given set of data was also included as a feature, calculated using the linear least squares method, as was the sum of the squares of

the residuals. To check for trend variations in the data that may have otherwise been hidden, a unique 'split-slope' approach was used wherein the same two features were extracted using just the first and second half of each series. Comparing the three slopes and residual values allowed for a better understanding of the underlying shape of the data, and whether the individual experienced a significant change during the assessments. The residuals, coupled with the number of local extrema, offer a measure of both the signal's variability and jitter. The participant's sex, age, and height were also included in the extracted features.

Once these features had been generated for each signal from each subject for both phases 1 and 2, the difference between the two was calculated (phase 2 minus phase 1). This produced one dataset that described the change in values between the first and second phases of the experiment. The result was then normalized by dividing each number by the sum of the absolute values of the two phases, e.g. for $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$ the feature matrices from phases 1 and 2, respectively, the final dataset $\boldsymbol{P_f}$ was calculated as:

$$\boldsymbol{P_f} = (\boldsymbol{P_2} - \boldsymbol{P_1}) \oslash (|\boldsymbol{P_1}| + |\boldsymbol{P_2}|) \tag{3.3}$$

where $\oslash$ is the Hadamard matrix division. In cases where the sum in the denominator would be zero (e.g. the minimum of a signal was zero for both phases), it was instead replaced by one. As all of the signals were nonnegative, this normalization resulted in values ranging from -1 to 1, inclusive.

3.2.3 Label Assignment Based on Prescribed Medications

In addition to the RMR and VO_2 peak features, each subject's medication history was used in this study. For each participant, a list was generated consisting of the medication, its type, the dosage, the frequency of the dosage, and the start and stop dates, if applicable. Medications that had been prescribed before the start of the

intervention were indicated by a blank start date, and the end dates of those drugs still taken at the end of the study were likewise left empty. This allowed for the calculation of the number of medications taken by any individual both at the start of the study (phase 1) as well as the end (phase 3).

A label was given to each subject based on his or her medication history. As one aim of this study was to develop a method for predicting successful outcomes from the lifestyle intervention, each included individual had to have finished the intervention with complete test data. If a participant ended the study on more medications than initially prescribed, he or she was labeled a 'failure', and those that finished the two-year program taking fewer medications were considered a 'success'. Those that finished the study taking the same number of medications as they were at the start were considered neutral and were omitted. For the purposes of this study, only medications that were prescribed for controlling co-morbid weight related conditions (e.g. hypertension, type 2 diabetes, gastroesophagael reflux disease) were included. For example, allergy medications were ignored as the underlying condition is not affected by weight. Some medications (e.g. HMG Co-A reductase inhibitors 'statins') were also ignored in subjects with diabetes, as it is best practice to continue an individual on the drug irrespective of weight loss. In addition, dosage was not considered: while a reduction in prescribed dose could be viewed as a positive, the focus of this chapter includes a stricter definition of success. A second analysis was performed that included all medications in the label generation process, and the results are discussed below. When comparing these labels to the weight-based ones used in Chapter II, only 58% of them remained unchanged.

Forty-two subjects (51 ± 10 years, 50% female) were included in this study. Based upon medications taken at phases 1 and 3, 23 were labeled as successes while the remaining 19 were categorized as failures. The initial dataset included 283 features per subject, extracted as described above from the physiological time-series data

collected during the RMR and VO_2 peak assessments from phases 1 and 2.

3.2.4 Predictive Model Generation Using Machine Learning

Machine learning algorithms were then applied to the dataset using the Waikato Environment for Knowledge Analysis (Weka); for each test, 10-fold cross-validation was used [71]. In this approach, the data is first split randomly into ten chunks. The algorithm is then run ten times, with each run using a different subset for testing and the remaining nine for training. The results of each of the runs are then averaged together to produce the final accuracy. For the Support Vector Machine (SVM), Random Tree, Neural Net, and Random Forest algorithms this total process was done ten times, with a different starting seed value used each run. The total accuracies, specificities, and sensitivities were then averaged. In the case of the SVM algorithm, the starting seed made no difference on the final results. For the K-Nearest Neighbor method, the algorithm was run using one, two, three, and four as values for K, with the total results again averaged. The Naïve Bayes algorithm was run only once, as there was no starting seed value parameter.

After the initial examination, feature selection was performed to reduce the number of features and remove unwanted noise. Due to the limited size of the dataset, this process was done using the entirety of the available input; to reduce the impact of any resultant bias, 10-fold cross-validation was used. The features were evaluated using information gain, a method for determining the usefulness of a feature in distinguishing between classes. For a given set of training instances, S, each entry with n features and a label r, the information gain of feature f is:

$$I_g(S, f) = H(S) - \sum_{v \in F} \frac{|\{\mathbf{s} \in S | s_f = v\}|}{|S|} \cdot H(\{\mathbf{s} \in S | s_f = v\}) \tag{3.4}$$

where $H(y)$ denotes the entropy of y, F the possible values for feature f, and s_f the value of feature f in instance $\mathbf{s}$. The resulting value, or merit, is a measure of how

much predictive power that feature has in modeling the label. As the information gain method was run ten times on different subsets of the dataset, each returning potentially different results, the average output for each feature was taken across all runs.

Table 3.3: Average information gain merit for the 25 most significant features; those with the same asterisk counts are duplicates

Feature	**Average Merit**
Number of turns, VO_2 (RMR test)*	0.295 ± 0.044
Number of turns, METS (RMR test)*	0.295 ± 0.044
Number of turns, VO_2\kg (RMR test)*	0.295 ± 0.044
Sex	0.086 ± 0.029
Second-half fit slope, respiratory rate (VO_2 test)	0.067 ± 0.103
Second-half fit residuals, respiratory rate (VO_2 test)	0.023 ± 0.069
Max of treadmill speed (VO_2 test)**	0.111 ± 0.14
Range of treadmill speed (VO_2 test)**	0.111 ± 0.14
Second-half fit slope, VCO_2 (VO_2 test)	0.106 ± 0.106
Number of turns, resting energy expenditure (RMR test)	0.12 ± 0.063
Second-half fit residuals, treadmill speed (VO_2 test)	0.212 ± 0.112
75^{th} percentile, VCO_2 (VO_2 test)	0.021 ± 0.064
Full fit residuals, VCO_2 (VO_2 test)	0.025 ± 0.074
Full fit slope, treadmill speed (VO_2 test)	0.088 ± 0.108
Second-half fit residuals, treadmill elevation (VO_2 test)	0.021 ± 0.064
First-half fit residuals, respiratory quotient (RMR test)	0.049 ± 0.099
Range of resting energy expenditure (RMR test)	0.021 ± 0.063
Range of respiratory quotient (RMR test)	0.029 ± 0.088
Range of VO_2 (RMR test)	0.021 ± 0.063
Max of F_ECO_2 (RMR test)	0.021 ± 0.063
Range of METS (RMR test)	0.021 ± 0.063
75^{th} percentile, VO_2\kg (RMR test)	0.026 ± 0.079
75^{th} percentile, METS (RMR test)	0.026 ± 0.079
Range of VO_2\kg (RMR test)	0.021 ± 0.063
Second-half slope, treadmill speed (VO_2 test)	0.023 ± 0.069

This evaluation complete, the top 25 features were selected and copied into a separate file. As can be seen in Table 3.3, a number of the selected features were identical. These were pared down to include only unique columns, resulting in 22 useful attributes. The columns containing the max and range of the treadmill speed were identical for all but one subject: in that case the difference was negligible and likely due to recording error (the speed should always start at zero, so the range should be equal to the max), so the columns were treated as equal. The same machine learning algorithms were applied, as above, to the newly-pared data, and the results were significantly improved. As before, the starting seed made no impact on the results of the SVM algorithm.

3.3 Results

The SVM machine learning algorithm was used to develop a preliminary predictive model based on the input features, and it outperformed other classification methods (Table 3.4).

Table 3.4: Results of various machine learning algorithms on the unreduced dataset

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Score**
Naïve Bayes	40.5%	0.39	0.42	0.42
Random Forest	52.1 ± 7.8%	0.60 ± 0.08	0.43 ± 0.10	0.58 ± 0.07
Random Tree	52.6 ± 8.1%	0.59 ± 0.07	0.43 ± 0.15	0.58 ± 0.07
K-Nearest Neighbor	56.0 ± 7.4%	0.50 ± 0.20	0.63 ± 0.15	0.54 ± 0.14
Neural Net	61.9 ± 1.1%	0.61 ± 0.00	0.63 ± 0.02	0.64 ± 0.01
SVM	**64.3 ± 0.0%**	0.78 ± 0.00	0.47 ± 0.00	0.71 ± 0.00

After the feature set was reduced using information gain, another SVM model was generated with the smaller dataset. In this scenario the Naïve Bayes algorithm slightly outperformed the SVM method in terms of both accuracy and combined sensitivity and specificity (Table 3.5). As this method relies on products of probabilities, having

a large number of noisy features can greatly affect the results [72]. While the Naïve Bayes approach assumes independence between features, given the label, this is not a valid assumption for this particular setting. A number of characteristics are highly correlated, with some being linear transforms of others (e.g. the VO_2/kg signal is simply the VO_2 series divided by the particular individual's weight, meaning a number of the extracted features are similarly related). As a result of this faulty assumption, the steps used to reduce and simplify the underlying probabilities are invalid and the predictive power of the generated classifier is compromised. By reducing the feature set to only relevant and independent entries, the observed significant jump in accuracy is not unexpected. Indeed, by selecting only those data that contribute meaningfully to generating a predictive model, the accuracies increase across the board. While not the top performer, the SVM model still returned markedly improved results over its implementation on the larger dataset and was only marginally behind the Naïve Bayes model. This latter method also saw such an improved score due to its formulation: by performing feature selection and reduction the number of confounding input vectors is limited.

Table 3.5: Results of various machine learning algorithms on the reduced dataset

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Score**
Random Tree	65.9 ± 5.7%	0.70 ± 0.10	0.62 ± 0.08	0.69 ± 0.06
K-Nearest Neighbor	70.8 ± 6.0%	0.70 ± 0.12	0.72 ± 0.05	0.72 ± 0.08
Neural Net	76.7 ± 2.2%	0.75 ± 0.04	0.78 ± 0.02	0.78 ± 0.02
Random Forest	77.4 ± 2.3%	0.78 ± 0.04	0.76 ± 0.04	0.79 ± 0.02
SVM	83.3 ± 0.0%	0.83 ± 0.00	0.84 ± 0.00	0.84 ± 0.00
Naïve Bayes	**85.7%**	0.83	0.90	0.86

3.4 Discussion

Based on these results it is clear that the normalized change in a select few features extracted from RMR and VO_2 peak tests, taken before and after a weight-loss intervention, can predict an individual's likelihood to reduce his or her number of medications at two years with over 85% accuracy. While RMR and VO_2 peak assessments were used in this study, the simple fact that an accurate prediction of an individual's two-year prognosis can be made using only data collected at baseline and after rapid weight loss has incredible clinical and financial implications for the healthcare industry. For physicians, knowing which patients will maintain the benefits of rapid weight loss, as measured by the number of prescribed medications, would prove useful in managing their long-term care. For example, if two individuals present with the same initial results after intense dieting, but one is classified as likely to be on more medications in the future, the clinician could choose different courses of action for each person. For 'successes', maintaining the current treatment plan would be appropriate; for the 'failures' an alternative approach could be implemented (e.g. bariatric surgery), potentially saving both doctor and patient over a year of time and effort. This approach could be applied prior to an intervention to guide clinicians in determining the strategy and/or level of intensity of treatment to offer.

In addition to the patient side of the clinical implications, the significance of the research side should also be considered. The results presented above are indicative of a positive, underlying physiological change induced by weight loss that persists regardless of any weight regain. In looking at the selected features (Table 3.3), a number of intriguing observations can be made. For example, the RMR and VO_2 peak tests are nearly evenly represented: after removing duplicates, the former accounts for 11 features and the latter 10. This shows an equal importance of both an individual's energy spent at rest and his or her level of fitness. In addition, there is a similarly even split between the statistical and algebraic features. Interestingly, the selected

characteristics derived from the RMR test tend to fall into the first group while the second group is composed primarily of features extracted from VO_2 peak test signals. This implies that, while both types of features are important, the overall changes observed in the resting data are more significant than how those changes progress. The opposite is true of the exercise data: the slope and best-fit residuals are generally more valuable than the absolute measurements. One potential explanation of this phenomenon is that an individual's RMR is fairly static, meaning measurements related to its value are more useful than those related to its short-term movement. While his or her VO_2 peak is also not subject to rapid fluctuations, the process whereby an individual reaches it is significant.

On a larger scale, previous randomized controlled clinical studies have demonstrated that weight loss can reduce the risk of progression to incident disease or reduce the number of medications used to treat established disease [73, 74]. This study takes that research and builds on it, showing that the persistence of these reductions in medications can be determined over a year ahead of time. Because a number of the medications taken by various participants directly affect an individual's weight, some clinicians consider a reduction in medication a more clinically relevant outcome than weight loss. While an individual's long-term health, and therefore medication regimen, can be affected by weight loss, one surprising result of this analysis was that this is mostly independent of any future weight regain. In fact, if the labels are defined strictly by weight loss as in Chapter II (a certain percent lost and kept off between phases 1 and 3) instead of by a persistent reduction in prescribed medications, the results are vastly different. Conventional machine learning algorithms produce no useful results, and including other input does not affect this low accuracy. Both time and frequency characteristics were included through the extraction of features from Fourier and wavelet transforms, but no meaningful models could be created. These poor results may be related to the fact that only approximately two-thirds of

individuals remain the same class when the criteria are changed (i.e. one third of individuals switch from 'success' to 'failure' or vice versa). Such a high incidence of unstable labels suggests a minimally causal relationship.

3.4.1 Future Work

In future studies, additional more advanced machine learning algorithms could be employed to further increase the classification accuracy. Other features should also be considered, including some related to the Fourier Transforms of the input signals. The labels themselves will also be explored more thoroughly, with dosages and types of medications being considered. While individuals that saw neither an increase nor a decrease in their number of prescribed medications were excluded from this study, future research could examine this group in a number of different ways. For example, by including dosages one could define 'success' as those on lower or more infrequent dosages. Alternatively, this relatively stable population could be considered a third and separate class, further refining the scale and allowing for borderline cases to be examined more closely or using alternative methods.

3.5 Summary

Overweight and obese adults constitute over two-thirds of the American population. Many are prescribed numerous medications to treat obesity-related diseases. To be able to accurately predict which individuals will require fewer medications after a 3-6 month lifestyle intervention could revolutionize the way in which obesity is managed. This chapter presents a method whereby the number of prescribed weight-related medications in an individual's regimen can be predicted to increase or decrease with an accuracy of nearly 85%. While this has important implications for health care, industry, and insurance providers, future studies will expand upon these results to improve their accuracy and resolution.

CHAPTER IV

Laplacian of Correlation Graph Classification: A Graph-Based Approach to Analyzing Noisy Datasets [1]

4.1 Background

As medical technologies progress and monitoring devices grow more intricate, the number and resolution of collected data increases. Coupled with the reduction in cost of high-throughput molecular and cellular measurement systems, this results in a wide variety of information available for any given patient [76, 77]. While this diversity can prove beneficial in uncovering otherwise subtle issues, the analysis of such large and disparate amounts of data can be challenging. Machine learning, a process whereby a dataset is fed into an algorithm that attempts to build a classification model for future data, is a promising solution. When coupled with feature extraction - choosing only the most useful datapoints - these methods can prove quite effective [78, 79]. However, the task of selecting which data to include can be daunting, and the inherent noise in the individual measurements must be considered [80]. While numerous methods exist for analyzing clean, curated data, these approaches are often plagued by poor accuracy when confronted with noisy datasets. A number

[1] Sections of this chapter were previously submitted as Biwer et al. [75].

of machine-learning algorithms perform some "built-in" feature selection, though this process can be time-consuming [81]. This chapter presents a graph-based approach to machine learning, Laplacian of Correlation Graph classification (LCG), that can outperform current approaches when applied to a noisy, unreduced dataset.

4.2 Methods

4.2.1 Data

The dataset used in this study is constructed in the same way as in Chapter III. To begin, signals are collected during RMR and VO_2 peak tests before and after an intensive lifestyle intervention (phases 1 and 2, respectively). Each individual's medication list is also noted. These are measured again two years after the initial visit (phase 3) in order to determine the long-term effects of the diet. To generate the set of features used for this research, characteristic values are computed from each signal collected during the first two RMR and VO_2 tests. The difference between the two sets is taken (phase 2 minus phase 1), and then each feature is normalized by the sum of its absolute values from the same tests (|phase 1| plus |phase 2|). This results in a dataset with values ranging from -1 to 1, inclusive.

In addition to these features, a label of 'success' or 'failure' is calculated for each participant based on his or her medication history. For an individual to be considered a 'success', he or she had to be taking fewer medications at the conclusion of the study (phase 3) than at baseline (phase 1). Conversely, a subject is labeled a 'failure' if he or she concluded the study taking more medications than before dieting. Any participant that was taking the same number of medications at phase 3 as phase 1 was excluded from the study; by eliminating these borderline cases, only the extremes were left to be considered. The prescribed dosage was not taken into consideration as a means to strengthen the validity of the labels. While a reduction could be considered positive

(or an increase considered negative), looking at each medication as a binary variable results in a more rigorous definition of 'success'. It should be noted that not all medications were included in the generation of these labels: only those prescribed for managing weight-related co-morbid conditions (e.g. gastroesophagael reflux disease, hypertension, type 2 diabetes) were used. The reason for this reduction is to allow for a better insight into the effectiveness of dieting on related co-morbidities. If allergy medications (or other non-weight-related medications) were included, they could skew the results away from the effects of the low-calorie lifestyle intervention towards other non-controlled changes. Additionally, some weight-related medications were also omitted for participants with diabetes as it is best practice to maintain the drugs despite any decrease in weight (e.g. statins).

This study includes data from forty-two subjects (50% female; average age 51 $\pm$ 10 years). Using the above method to generate labels, 23 are considered 'successes' while the remaining 19 are classified as 'failures'. The initial dataset is comprised of 280 features per subject, as described above.

4.2.2 Laplacian of Correlation Graph Classification

An overview of LCG is shown in Figure 4.1. With the dataset generated, the first step to LCG is to split the input into a training set and a testing set. Based on a preliminary grid search, this dataset was split such that there are 18 instances in the training group and the remaining 24 are used for testing. The search involved starting with two entries in the training set: one 'success' and one 'failure'. For each iteration the size was increased by two, with the 50/50 split between classes maintained. This was repeated until half of the available data was in the training set (20 subjects). The instance with 18 training and 24 testing cases yielded the best results across all runs. While the 18 entries in the training set are split evenly between 'successes' and 'failures', the exact combination of subjects was left to vary in the analysis and will

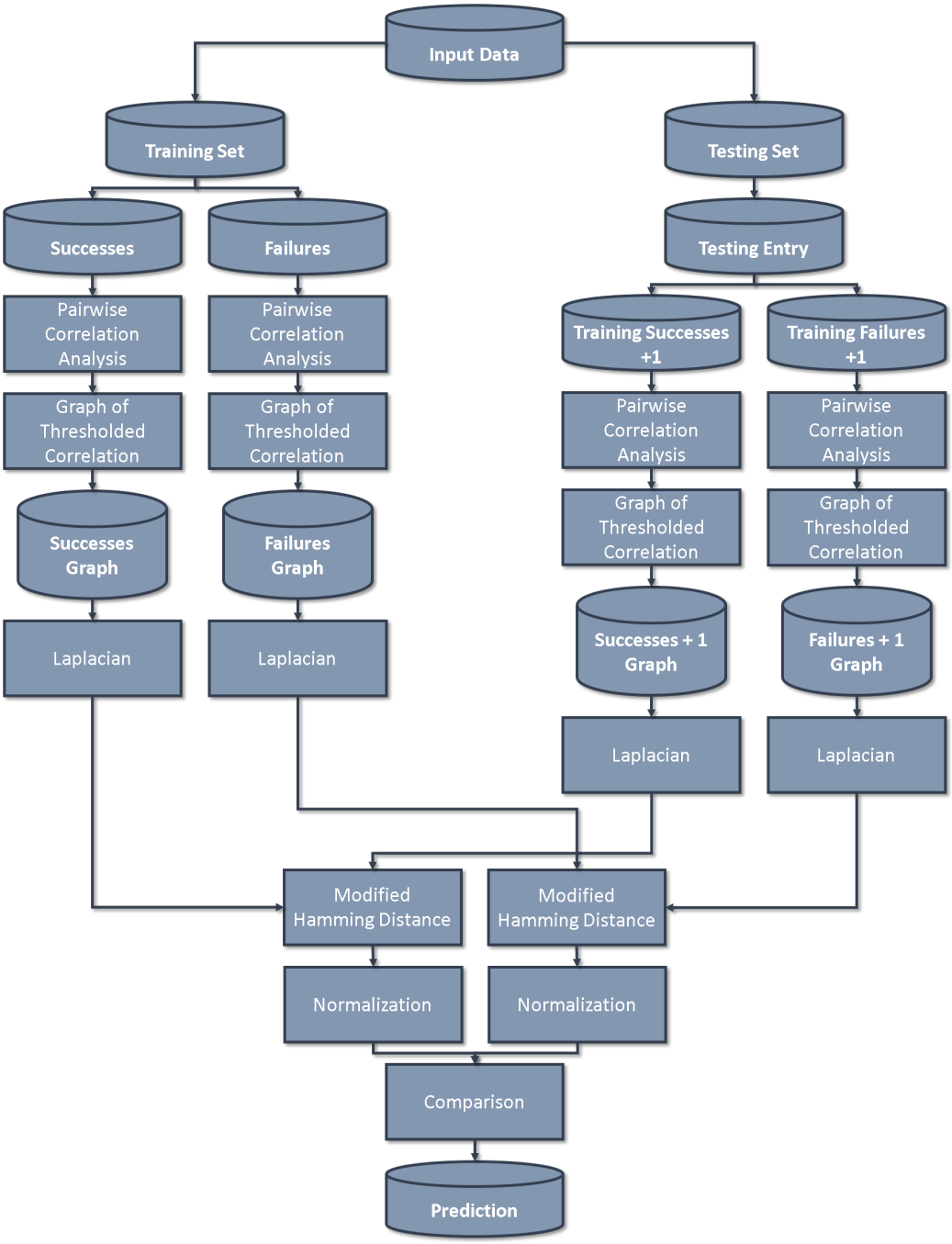

Figure 4.1: Schematic diagram of LCG

be further discussed below. Once this separation is decided, a matrix of features is generated for each of the two training groups ('success' and 'failure') where each row represents a unique participant and each column a feature. A graph is also generated for each training group such that each node represents a feature. For each pair of distinct nodes in the same graph, the two are connected if the correlation between the corresponding columns in the feature matrix has a t-test p-value below some set threshold h_p. For this study h_p was set to 0.31, again optimized through a grid search: as the value of a p-value is bounded by zero and one, trial values of h_p started at 0.01 and increased to 1.0 in steps of 0.01. The results of this search showed a value of 0.31 to yield the most consistent results. This complete, an adjacency matrix is created for each of the 'success' and 'failure' graphs: S_0 and F_0, respectively. These serve as the baseline from the training set.

Once these initial matrices are created, one instance T from the testing set is chosen and appended to the end of the two training matrices. The process is repeated as above, with the end result being an additional 'success' and an additional 'failure' graph, S_t and F_t. Using a modified version of the Hamming distance, the difference S_d between the Laplacian of S_0 and that of S_t is calculated, as is that between the Laplacians of F_0 and F_t (denoted F_d). The Laplacian $\boldsymbol{\mathcal{L}}$ of a matrix is the difference between its degree and adjacency matrices, or:

$$\boldsymbol{\mathcal{L}} = \boldsymbol{G} - \boldsymbol{J} \tag{4.1}$$

where $\boldsymbol{G}$ and $\boldsymbol{J}$ are the degree and adjacency matrices, respectively [82]. The Laplacian was chosen because it contains information relevant to the related graph, e.g. each node's connectedness in terms of both number of neighbors as well as which particular vertices are nearby. By calculating the differences between Laplacian matrices, changes from one to the next become more exaggerated and any deviations from the baseline are highlighted.

To understand the modified Hamming distance used in this research, let A and B be any two $m \times n$ matrices. First define $N_{i,j}$ and $D_{i,j}$ as:

$$N_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} \neq 0 \text{ or } B_{i,j} \neq 0, \text{ but not both;} \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

and

$$D_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} \neq 0 \text{ or } B_{i,j} \neq 0; \\ 0 & \text{if } A_{i,j} = B_{i,j} = 0. \end{cases} \tag{4.3}$$

With these, the modified Hamming distance can be calculated as:

$$\text{modHam}(A, B) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} N_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} D_{i,j}} \tag{4.4}$$

As such, define S_d and F_d as:

$$S_d = \text{modHam}(\mathcal{L}(S_0), \mathcal{L}(S_t)) \tag{4.5}$$

$$F_d = \text{modHam}(\mathcal{L}(F_0), \mathcal{L}(F_t)) \tag{4.6}$$

These distances are then normalized by the number of nonzero entries in the initial training-set graphs (yielding S_{dn} and F_{dn}), thereby allowing a change in a relatively sparse graph to account for more of a difference than the same change in a more dense graph. The ratio S_{dn}/F_{dn} is taken and thresholded: if the value is above the cutoff the instance T is labeled a 'success', and otherwise labeled a 'failure'. This threshold was varied throughout the study to generate receiver operating characteristic curves, as discussed below. Figure 4.2 shows two example thresholded correlation graphs, a 'success' (top) and a 'failure' (bottom) generated from a training set with 67 features. In this instance the 'success' graph is noticeably less dense: there are a total of 1316 connections between nodes, while there are 1620 edges in the 'failure' graph. Adding

in an entry from a testing set and regenerating the plots changes their densities. The new 'success' graph has 1418 connections, an increase of 92, while the new 'failure' graph only sees a decrease of 6 edges to 1614. The former not only sees a more marked change in connectedness, but because of its relative sparsity the normalized difference is even more pronounced. As such, the added entry from the testing set is labeled a 'failure', a correct classification.

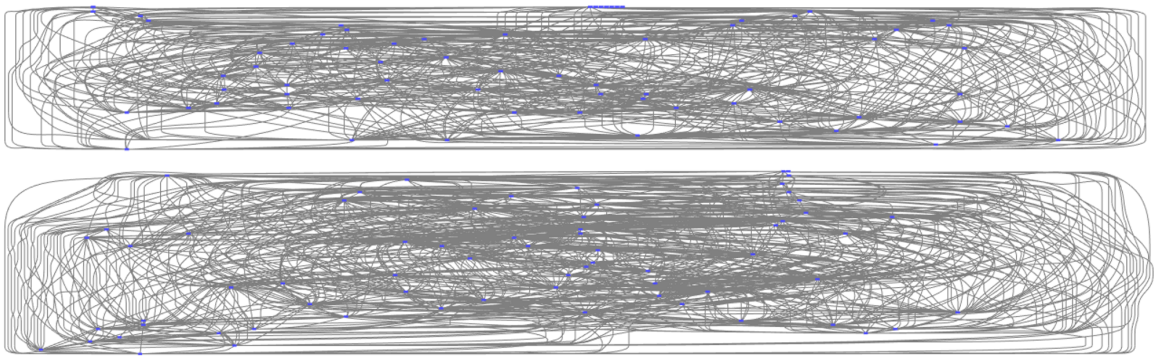

Figure 4.2: Example LCG Graphs

4.3 Results

To test the robustness of LCG, as well as its effectiveness at handling noise, numerous variations of the dataset were examined and compared the results to those of other well-established methods.

In the first case, the full 280 features described above were used. The rows were randomly permuted 100 times under the constraint that the first 18 rows had to be evenly split between 'successes' and 'failures'. In each permutation these initial rows were then used as the training set, with the remaining 24 instances comprising the test set. With this split defined LCG was applied to the dataset, each time generating a Receiver Operating Characteristic (ROC) curve - a measure of the robustness of a classifier - by varying the final ratio threshold. The Area Under the ROC Curve (AUC) is used to gauge how resilient a method's accuracy is to perturbations of its

discrimination threshold, with larger values preferred. In addition to LCG, WEKA was utilized to apply four standard machine learning approaches to the same datasets with the same training / testing splits: Naïve Bayes, support vector machine (SVM), random forest, and random tree. The results are shown in Table 4.1.

Table 4.1: Results with 280 features

Method	**Average Accuracy**	**Average AUC**
SVM	50.4%	.505
Random Forest	51.0%	.518
Naïve Bayes	51.2%	.506
Random Tree	52.1%	.521
LCG	**58.2%**	**.568**

After generating these results, the size of the dataset was reduced to 197 columns by eliminating a selection of noisy features. As described in Chapter III, this was done by calculating the 'information gain' of each column, a measure of the information inherent within any given feature. This process was repeated twice more, generating two additional datasets with 137 and 67 features. Appendix A lists which features were included at each step. Each of these three feature sets was randomly permuted 100 times, again with the stipulation that the first 18 rows be split evenly between the two classes. This complete, LCG and the other four prominent machine learning algorithms were applied to the resulting datasets; the results are shown in Table 4.2.

Looking at this table, it is easy to see the extent to which LCG outperforms the other standard approaches, regardless of the level of noise inherent in the dataset. Indeed, even in the most difficult case LCG scores an average accuracy above 58%; while in itself not particularly impressive, the fact this represents more than an 11% increase over the next-best method lends credence to the effectiveness of the graph-based approach.

If the average AUC values are examined, it is even more clear that LCG presents an improvement over standard approaches when applied to noisy datasets. Figure 4.3 shows a selection of ROC curves typical to each method for the smaller two datasets,

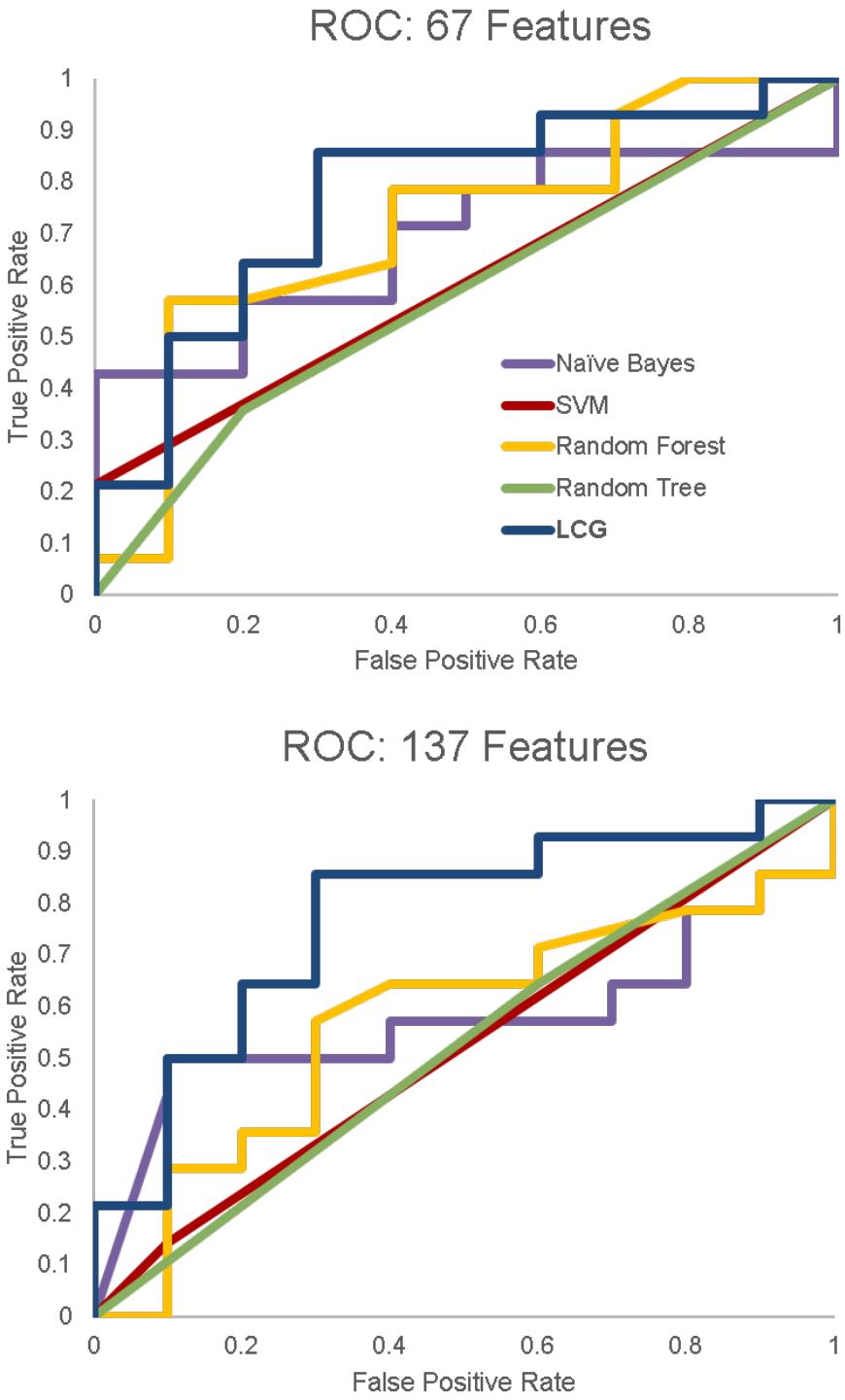

Figure 4.3: Typical ROC graphs for various dataset sizes

Table 4.2: Results with smaller feature sets

Method	**Average Accuracy**	**Average AUC**
	197 Features	
SVM	51.7%	.511
Random Forest	55.2%	.565
Naïve Bayes	54.2%	.557
Random Tree	52.0%	.515
LCG	**59.9%**	**.639**
	137 Features	
SVM	50.5%	.525
Random Forest	55.0%	.579
Naïve Bayes	55.1%	.576
Random Tree	52.7%	.527
LCG	**59.3%**	**.650**
	67 Features	
SVM	61.2%	.614
Random Forest	63.7%	.712
Naïve Bayes	65.9%	.700
Random Tree	58.2%	.584
LCG	**66.0%**	**.774**

further illustrating the fact that LCG is generally more robust. The ROC plot for the dataset with 137 features, in particular, shows a significant advantage for LCG. When the AUC data from Table 4.2 are plotted (Figure 4.4), they show a clear trend shared between the five methods. The values for LCG, however, are consistently higher than the standard approaches. This bias is further evidence that the graph-based approach outperforms other algorithms on noisy datasets.

4.4 Discussion

As is shown in the above results, LCG outperforms other standard machine learning methods over a wide array of noisy datasets. While there are a number of variables that need to be tuned to best optimize LCG, these can be easily determined by a brief grid search. Indeed, the customization options allow for LCG to be applied to a wide array of problems, not just those in the medical field. For example, if a case

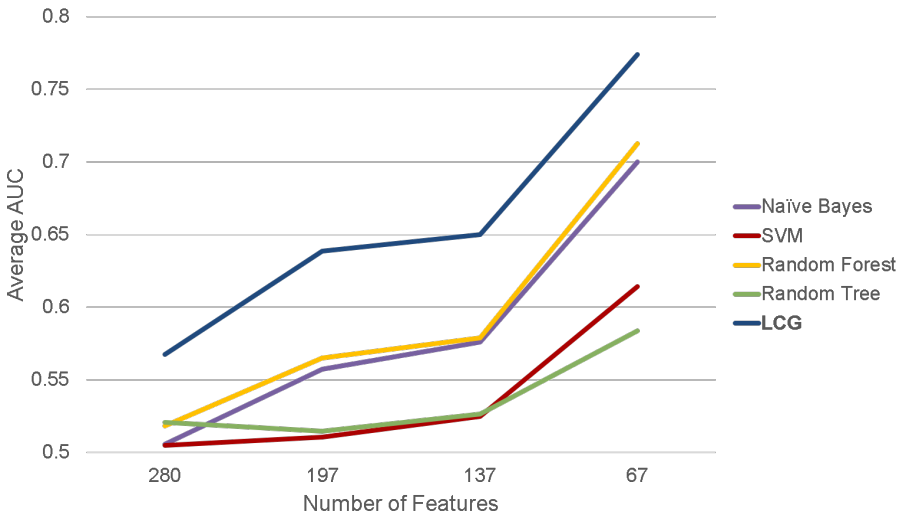

Figure 4.4: Average AUC values for various dataset sizes

were to present itself in which one class was heavily favored over another, the final ratio threshold could be altered so as to weight the classification towards the more likely group. Another alternative would be to skew the size of the training set: if one class has more training entries than another, it would take a more drastic change from the testing input to overcome this bias. Like the SVM algorithm, the underlying kernel in LCG could even be changed: instead of using a modified Hamming distance, other similarity functions can be put in its place. This allows for more complex and tailored approaches to determining the likeness of any two given connected graphs. Extending this concept, it would be easy to convert LCG from a binary to a multi-class classifier. After generating a training graph for each class, dissimilarity scores can be calculated for a given test case by adding it to each. Instead of calculating and thresholding a ratio as above, any method could be used to select to which class the example belongs (e.g. the class with the smallest dissimilarity score).

One other advantage to LCG is the ease and extent to which it can be parallelized. With the prevalence of multi-core machines and cloud computing, the ability of an algorithm to make the most of distributed calculations can result in huge decreases in runtime. The creation of the training graphs can easily be split, even to the

point of parsing out node-pairs to determine if they are connected. Once these initial adjacency matrices are calculated, each testing case can be run independently (and each can be further split by again parsing out node-pairs).

While the ROC curves show LCG to be robust, because of the large number of steps involved changing a parameter at the beginning of the process (e.g. the size of the training set, or the threshold for joining two nodes) can greatly affect the final results. This can easily be addressed by performing a quick grid search to determine what values will work best for a given problem, but it means that there are not any 'out-of-the-box' parameter values that are guaranteed to give better-than-average results.

4.4.1 Future Work

In future studies, a more lenient approach to generating the class labels could be used (e.g. dosage could be considered). Alternatively, the training set could be chosen from the current dataset, with the testing set built to include participants with more loosely-defined labels. As was mentioned in Chapter III, other features could also be included. Characteristics derived from Fourier transforms may prove useful, and wavelet transforms could allow for relationships between time and frequency to help differentiate the two classes. As to improving the algorithm, other kernels could be used to generate different similarity scores between graphs. Missing data could also be incorporated into the dataset, and the approach could be altered to weigh certain features over others accordingly. Likewise, instead of thresholding the edges (thus setting them to 0's and 1's), the p-values could be left as weights. This would however also necessitate a more intricate method for determining the similarity of any two graphs, as the current implementation takes advantage of the binary nature of the edges.

4.5 Summary

Overweight and obesity are major problems currently facing the healthcare industry. With over two-thirds of the US population overweight, the burden on both physicians and patients is significant. This strain is readily apparent when looking at the number of medications many overweight individuals are prescribed to help manage their weight-related co-morbidities. While many people attempt weight-loss as a means to improve their health and reduce the size of their medication regimen, there are as of yet no ways to predict whether that weight-loss will translate to fewer pills in the long term. While machine learning is ideally suited to this problem, with ever-improving medical monitoring technologies and the ready availability of digital storage the feature space can be dauntingly large. This chapter not only shows a method for predicting the relative size of an individual's medication regimen, but also demonstrates that LCG is better-suited to noisy datasets. By allowing for less-clean input at a relatively small cost to accuracy LCG has the potential to impact a wide array of fields, not just those related to healthcare.

CHAPTER V

Contributions and Insights

In this chapter, the implications of the research are examined both in terms of their effects on the relevant fields and also in the underlying ramifications of the presented results. By studying the windowed persistent homology method described in Chapter II, a clear difference can be shown between those for whom dieting will prove effective and those for whom it will not. Applicable to any time-series, this signal processing algorithm can be used to uncover subtle differences in problems outside the clinical field as well. Chapter III presents a means other than weight loss by which the efficacy of dieting can be measured, which in turn allows for a number of interesting observations to be made pertaining to an individual's overall health. Finally, the machine learning algorithm developed in Chapter IV shows promise in mining noisy datasets. This graph-based approach can be easily distributed across multiple processors, and the results are amenable to interpretation.

5.1 Signal Processing for Analyzing Activity Data

5.1.1 Contributions of Windowed Persistent Homology

In studying overweight and obesity, the steady improvements in monitoring technologies have allowed for a growth in available signal data. While current methodologies for processing time series are readily applicable to this new dataset, they are not

always effective. For example, standard signal processing algorithms yield no useful results when applied to the activity data presented in Chapter II. By developing a windowed variation of persistent homology, along with a modified version of the Hausdorff distance, it is possible to detect a difference between individuals likely to lose weight through dieting and maintain the loss and those for whom caloric restriction will be less effective. While future studies are needed to investigate the intricacies of persistence diagrams as they pertain to movement profiles, this research has shown a clear separation where other methods failed to do so. Through the inclusion of more data, a predictive model can be built with the potential to save the time, effort, and money of clinicians and patients alike.

5.1.2 Insights Gained by Analyzing Activity Data

These promising results lend themselves to additional questions and avenues of research. For example, the role of the signal's sampling frequency must be considered: in the above study, the armband data was aggregated and recorded once every minute. By increasing or decreasing this resolution, certain types of movement may become more apparent while those present in the current movement profiles are obscured. As these different activities shift in and out of focus, the split between the two groups may grow or shrink. Another question to consider is whether or not an individual can alter his or her lifestyle enough to effectively switch groups. By changing one's behavior to better align with the movement profiles of those labeled 'successes', it may be possible to increase the long-term efficacy of dieting. Perhaps the most promising implication of this research is that it may one day be possible to predict this efficacy before even attempting any caloric restrictions.

5.2 Machine Learning Applied to Physiological Data to Predict Future Prescription Medication Use

5.2.1 Contributions to Defining Weight Loss Success

While weight loss is an obvious metric for determining the success of dieting, other factors must also be considered. For example, long-term maintenance of any loss should be noted, as many that succeed in losing weight end up regaining some or all of their loss. While this can be viewed as negating the original success, there may be other underlying benefits that persist through the regain: these can be thought of as successes in their own right. By altering the metric by which an individual's progress is evaluated, the results of any subsequent analyses may be markedly different. For example, utilizing the same dataset presented in Chapter III but with labels generated by thresholding initial weight loss and long-term weight-loss maintenance, the demonstrated machine learning algorithms produce no significantly useful models. By instead focusing on medication regimens, a simple, systematic definition of success can be created. As shown above, looking at the number of weight-related prescription medications makes it possible to generate meaningful class labels. This approach differs from current methodologies in that it values relative change over absolute thresholds, allowing for a more personalized experience.

5.2.2 Insights Gained from Predictive Modeling

Despite the impressive accuracy of the predictive models created using these labels, a number of questions remain. One such issue pertains to the removal of medications deemed to be unrelated to weight: if ignoring these drugs alters the results, their influences must be further studied. Certain medications in particular were excluded for diabetic subjects during label generation because of clinical best-practice procedures; by examining those with diabetes separately from those without the condition,

it would be possible to control for this disparate treatment. While this would require a more substantial dataset, additional subjects could be included if a means were developed to classify those previously unlabeled. These edge-case individuals who saw neither an increase nor a decrease in their medication regimen may hold valuable information pertaining to the role weight actually plays in overall health. Taken together, the three groups may allow clinicians a better understanding of how an individual can regain lost weight but not previously-diagnosed weight-related complications. This understanding may be furthered by also examining the selected features. Both the RMR and VO_2 peak tests produced useful signals, but the underlying reasons for their significance should be considered. The statistical nature of the resting test's notable characteristics indicates the value lies in the absolute measurements, while the more complex algebraic features typically chosen from the exercise test signify the manner in which the data change is more informative.

5.3 Laplacian of Correlation Graph Classification

5.3.1 Contributions to Graph-Based Machine Learning

As datasets continue to grow in size and complexity, so too must the methods used to analyze them. Feature selection is an integral part of this process, but it must be balanced against filtering out potentially informative data. With large inputs come the possibilities of intricate relationships between different aspects of the data, which means that while a particular feature may not be useful on its own it may prove valuable when paired with another. As such, removing data through feature selection may limit the potential effectiveness of a generated model. However, skipping this step will result in noisy datasets. Chapter IV introduces LCG, a graph-based method capable of outperforming a number of standard approaches when applied to a noisy feature set. By examining the pairwise correlations between features, important relationships

may be discovered that would have otherwise gone unnoticed. In addition, because each pairing is independent of the next, the calculations can be broadly distributed without any loss in accuracy.

5.3.2 Insights Gained on Processing Noisy Data

While LCG has proven useful in examining physiological signals as they relate to medication regimens, its potential applications are much more widespread. Other sources of data can be incorporated into the input, allowing for the discovery of previously unknown relationships between disparate measurements. Not limited to the medical field, LCG can also be used in any setting where machine learning would be applicable. In particular, situations that involve large amounts of noisy data are ideal candidates, as are classification problems involving more than two classes. In addition, the structure of the algorithm allows for it to be easily parallelized: by distributing the correlation calculations, the computation time required is limited only by the available processing power. The generation of graphs to represent the underlying feature set is also an advantage as it allows for easy visual comparisons between classes.

CHAPTER VI

Conclusion and Future Directions

6.1 Conclusion

With overweight and obesity on the rise, the health and financial impacts of their associated diseases and conditions will only continue to grow. Many individuals attempt volitional weight loss through diet and exercise, but no current methods exist that yield *a priori* knowledge of success. However, through the use of advances in clinical monitoring capabilities and modern signal processing methods, paired with innovative machine learning algorithms, models for predicting an individual's chances of success are becoming possible. By studying the movement profiles of overweight and obese individuals, a pattern emerges that shows a clear difference between those likely to lose weight and those for whom dieting will be less effective. This alone has major implications on the healthcare industry: not only does it allow for massive potential savings of time and resources through the avoidance of ineffective treatment options, but it also shows an underlying predisposition that can and should be further studied. In addition to long-term weight loss, a reduction in an individual's medication regimen is indicative of an improvement in his or her general health. Through features calculated from physiological signals that can be measured non-invasively, a model can be built that accurately predicts long-term medication use. This too has the potential to drastically affect the healthcare industry.

With the methods and algorithms presented in this thesis, predicting the efficacy of dieting on overweight and obesity is possible. Future studies are needed to refine and improve these approaches, but the investments will no doubt have a lasting impact. Both the medical field and the healthcare industry stand to benefit, as these diseases and their underlying causes and associated conditions are a major epidemic facing not only the United States but the entire world.

6.2 Future Directions

The work presented in this thesis is a step towards a better understanding of overweight and obesity. While the results are promising, more research is needed to improve not only the accuracy and robustness of the methods but also the medical understanding of the diseases. For example, the difference in movement profiles shown in Chapter II has the potential to drastically change the way in which dieting is approached. As it stands, however, there are not enough cases to build a predictive model. Recording more data would allow for more robust results: from strengthening the evidence of an underlying difference to generating a predictive model, the potential benefits are staggering. In addition to gathering more data, future studies could investigate the medical reasons behind the separate groups. By understanding what aspects of an individual's movement profile are responsible for his or her long-term weight loss success, new diet and exercise strategies could be developed to help those not predisposed to beneficial behavior. From a signal-processing perspective, persistent homology has a wide range of potential uses. As the method has no requirements pertaining to sampling rate or other properties, it can be applied to virtually any time-series data. One potential application is heart-rate signals: by applying the method to healthy input, a persistence diagram template can be generated. This complete, any future signals can be processed and compared to the standard. If a difference threshold between the two plots is reached, it may be indicative of a cardiac

issue.

As mentioned in Chapter III, the criteria for success were quite strict. Future studies can relax these requirements, taking into account other medications or prescribed dosages. Those individuals that saw neither an increase nor a decrease in their medication regimens should also be more closely examined, as these border cases could hold valuable insights into their underlying conditions. The various combinations of comorbidities should also be considered: as individuals with diabetes were treated differently than those without, a study involving only members from one group may shed light on how the associated conditions affect each other. Likewise, any health concerns present in both overweight and lean individuals could be studied to examine the effects of other influences (e.g. diet composition, activity) on medication levels.

As data storage becomes cheaper and monitoring devices more intricate, applications for the algorithm presented in Chapter IV become more varied and widespread. Future studies could look to improve its accuracy or computational time, whether through tweaks to the underlying method or by exploiting any potential shortcuts. Other difference metrics could be used to compare the resulting graphs, as this is a critical step in the algorithm and could produce drastically different results. Applying LCG to other datasets would also prove useful, especially large or noisy collections of features, as this would demonstrate its ability to function on unreduced input. In particular, generating a feature set that includes additional clinically relevant measurements could produce more accurate results and allow for the discovery of previously unrecognized correlations across disparate sources of data. For example, incorporating the data previously omitted from the above studies may yield useful insights. Integrating the psychological questionnaires, bloodwork results, and DEXA scans with the RMR and VO_2 peak tests, tensors can be created to further strengthen and elucidate any underlying connections. By examining these interactions, the relationships between an individual's weight and his or her mental and physical health

can be studied. This could have major implications on the ways in which overweight and obesity are treated, as well as on how the efficacy of dieting is defined.

APPENDIX

APPENDIX A

Feature Selection Using Information Gain

Table A.3 shows which features were kept as the dataset used in Chapter IV was reduced in size. In the first column, a description of the feature can be found formatted as *signal_test_feature*. Table A.1 describes the shorthand used to indicate from which signal the feature is derived, and Table A.2 the shorthand used to indicate the feature. To differentiate between signals recorded during the RMR and VO_2 peak tests, an indicator is included: *rmr* indicates the former and *ex* the latter. While the features kept in the reduced dataset described in Chapter III had non-zero merits (as calculated by the information gain method), the remainder of the feature set was trivial. Due to this, the characteristics removed in each reduction were chosen at random, under the condition they had zero merit.

Table A.1: Legend of signal shorthands found in A.3

Shorthand	**Signal**
feco2	F_ECO_2
feo2	F_EO_2
mets	METS
ree	REE
rq	RQ / RER
rr	RR
tm_elv	Treadmill Elevation
tm_spd	Treadmill Speed
vco2	VCO_2
ve	V_E
vo2	VO_2
vo2kg	VO_2/KG

Table A.2: Legend of feature shorthands found in A.3

Shorthand	**Description**
_25th	The 25^{th} percentile value
_50th	The 50^{th} percentile value (median)
_75th	The 75^{th} percentile value
_avg	The average value of the signal (mean)
_fit	The slope of the best-fit line
_fit_1	The slope of the best-fit line for the first half of the data
_fit_1_residuals	The residuals of the first-half best-fit line
_fit_2	The slope of the best-fit line for the second half of the data
_fit_2_residuals	The residuals of the second-half best-fit line
_fit_residuals	The residuals of the best-fit (R^2)
_max	The global maximum recorded during the test
_min	The global minimum recorded
_numTurns	The number of local extrema
_range	The difference between the maximum and minimum values

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
feco2_ex_25th	x	x	x	
feco2_ex_50th	x	x	x	
feco2_ex_75th	x	x	x	
feco2_ex_avg	x	x	x	
feco2_ex_fit	x	x	x	
feco2_ex_fit_1	x	x	x	
feco2_ex_fit_1_residuals	x	x	x	
feco2_ex_fit_2	x	x	x	
feco2_ex_fit_2_residuals	x	x	x	
feco2_ex_fit_residuals	x	x	x	
feco2_ex_max	x	x	x	
feco2_ex_min	x	x	x	
feco2_ex_numTurns	x	x	x	
feco2_ex_range	x	x	x	
feco2_rmr_25th	x			
feco2_rmr_50th	x			
feco2_rmr_75th	x			
feco2_rmr_avg	x			
feco2_rmr_fit	x			
feco2_rmr_fit_1	x			
feco2_rmr_fit_1_residuals	x			
feco2_rmr_fit_2	x			
feco2_rmr_fit_2_residuals	x			
feco2_rmr_fit_residuals	x			

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
feco2_rmr_max	x	x	x	x
feco2_rmr_min	x			
feco2_rmr_numTurns	x			
feco2_rmr_range	x			
feo2_ex_25th	x	x	x	x
feo2_ex_50th	x	x	x	x
feo2_ex_75th	x	x	x	x
feo2_ex_avg	x	x	x	x
feo2_ex_fit	x	x	x	x
feo2_ex_fit_1	x	x	x	x
feo2_ex_fit_1_residuals	x	x	x	x
feo2_ex_fit_2	x	x	x	x
feo2_ex_fit_2_residuals	x	x	x	x
feo2_ex_fit_residuals	x	x	x	x
feo2_ex_max	x	x	x	x
feo2_ex_min	x	x	x	x
feo2_ex_numTurns	x	x	x	x
feo2_ex_range	x	x	x	x
feo2_rmr_25th	x	x		
feo2_rmr_50th	x	x		
feo2_rmr_75th	x	x		
feo2_rmr_avg	x	x		
feo2_rmr_fit	x	x		
feo2_rmr_fit_1	x	x		

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
feo2_rmr_fit_1_residuals	x	x		
feo2_rmr_fit_2	x	x		
feo2_rmr_fit_2_residuals	x	x		
feo2_rmr_fit_residuals	x	x		
feo2_rmr_max	x	x		
feo2_rmr_min	x	x		
feo2_rmr_numTurns	x	x		
feo2_rmr_range	x	x		
mets_ex_25th	x	x	x	
mets_ex_50th	x	x	x	
mets_ex_75th	x	x	x	
mets_ex_avg	x	x	x	
mets_ex_fit	x	x	x	
mets_ex_fit_1	x	x	x	
mets_ex_fit_1_residuals	x	x	x	
mets_ex_fit_2	x	x	x	
mets_ex_fit_2_residuals	x	x	x	
mets_ex_fit_residuals	x	x	x	
mets_ex_max	x	x	x	
mets_ex_min	x	x	x	
mets_ex_numTurns	x	x	x	
mets_ex_range	x	x	x	
mets_rmr_25th	x			
mets_rmr_50th	x			

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
mets_rmr_75th	x	x	x	x
mets_rmr_avg	x			
mets_rmr_fit	x			
mets_rmr_fit_1	x			
mets_rmr_fit_1_residuals	x			
mets_rmr_fit_2	x			
mets_rmr_fit_2_residuals	x			
mets_rmr_fit_residuals	x			
mets_rmr_max	x			
mets_rmr_min	x			
mets_rmr_numTurns	x	x	x	x
mets_rmr_range	x	x	x	x
ree_rmr_25th	x	x		
ree_rmr_50th	x	x		
ree_rmr_75th	x	x		
ree_rmr_avg	x	x		
ree_rmr_fit	x			
ree_rmr_fit_1	x			
ree_rmr_fit_1_residuals	x			
ree_rmr_fit_2	x			
ree_rmr_fit_2_residuals	x			
ree_rmr_fit_residuals	x			
ree_rmr_max	x	x		
ree_rmr_min	x	x		

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
ree_rmr_numTurns	x	x	x	x
ree_rmr_range	x	x	x	x
rq_ex_25th	x			
rq_ex_50th	x			
rq_ex_75th	x			
rq_ex_avg	x			
rq_ex_fit	x			
rq_ex_fit_1	x			
rq_ex_fit_1_residuals	x			
rq_ex_fit_2	x			
rq_ex_fit_2_residuals	x			
rq_ex_fit_residuals	x			
rq_ex_max	x			
rq_ex_min	x			
rq_ex_numTurns	x			
rq_ex_range	x			
rq_rmr_25th	x	x		
rq_rmr_50th	x	x		
rq_rmr_75th	x			
rq_rmr_avg	x	x		
rq_rmr_fit	x			
rq_rmr_fit_1	x	x		
rq_rmr_fit_1_residuals	x	x	x	x
rq_rmr_fit_2	x			

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
rq_rmr_fit_2_residuals	x	x		
rq_rmr_fit_residuals	x	x		
rq_rmr_max	x	x		
rq_rmr_min	x	x		
rq_rmr_numTurns	x			
rq_rmr_range	x	x	x	x
rr_ex_25th	x	x	x	x
rr_ex_50th	x	x	x	x
rr_ex_75th	x	x	x	x
rr_ex_avg	x	x	x	x
rr_ex_fit	x	x	x	x
rr_ex_fit_1	x	x	x	x
rr_ex_fit_1_residuals	x	x	x	x
rr_ex_fit_2	x	x	x	x
rr_ex_fit_2_residuals	x	x	x	x
rr_ex_fit_residuals	x	x	x	x
rr_ex_max	x	x	x	x
rr_ex_min	x	x	x	x
rr_ex_numTurns	x	x	x	x
rr_ex_range	x	x	x	x
tm_elv_ex_25th	x	x		
tm_elv_ex_50th	x	x		
tm_elv_ex_75th	x	x		
tm_elv_ex_avg	x	x		

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
tm_elv_ex_fit	x	x		
tm_elv_ex_fit_1	x	x	x	
tm_elv_ex_fit_1_residuals	x	x	x	
tm_elv_ex_fit_2	x	x	x	
tm_elv_ex_fit_2_residuals	x	x	x	x
tm_elv_ex_fit_residuals	x	x	x	
tm_elv_ex_max	x	x		
tm_elv_ex_min	x	x		
tm_elv_ex_numTurns	x	x		
tm_elv_ex_range	x	x		
tm_spd_ex_25th	x	x	x	x
tm_spd_ex_50th	x	x	x	
tm_spd_ex_75th	x	x	x	
tm_spd_ex_avg	x	x	x	x
tm_spd_ex_fit	x	x	x	x
tm_spd_ex_fit_1	x	x		
tm_spd_ex_fit_1_residuals	x			
tm_spd_ex_fit_2	x	x	x	x
tm_spd_ex_fit_2_residuals	x	x	x	x
tm_spd_ex_fit_residuals	x	x		
tm_spd_ex_max	x	x	x	x
tm_spd_ex_min	x	x	x	x
tm_spd_ex_numTurns	x	x	x	
tm_spd_ex_range	x	x	x	x

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
vco2_ex_25th	x	x	x	x
vco2_ex_50th	x	x	x	
vco2_ex_75th	x	x	x	x
vco2_ex_avg	x	x	x	x
vco2_ex_fit	x	x	x	
vco2_ex_fit_1	x	x	x	
vco2_ex_fit_1_residuals	x	x	x	
vco2_ex_fit_2	x	x	x	x
vco2_ex_fit_2_residuals	x	x	x	
vco2_ex_fit_residuals	x	x	x	x
vco2_ex_max	x	x	x	x
vco2_ex_min	x	x	x	x
vco2_ex_numTurns	x	x	x	
vco2_ex_range	x	x	x	x
vco2_rmr_25th	x	x		
vco2_rmr_50th	x	x		
vco2_rmr_75th	x	x		
vco2_rmr_avg	x	x		
vco2_rmr_fit	x	x		
vco2_rmr_fit_1	x	x		
vco2_rmr_fit_1_residuals	x	x		
vco2_rmr_fit_2	x	x		
vco2_rmr_fit_2_residuals	x	x		
vco2_rmr_fit_residuals	x	x		

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
vco2_rmr_max	x	x		
vco2_rmr_min	x	x		
vco2_rmr_numTurns	x	x		
vco2_rmr_range	x	x		
ve_ex_25th	x	x	x	
ve_ex_50th	x	x	x	
ve_ex_75th	x	x	x	
ve_ex_avg	x	x	x	
ve_ex_fit	x	x	x	
ve_ex_fit_1	x	x	x	
ve_ex_fit_1_residuals	x	x	x	
ve_ex_fit_2	x	x	x	
ve_ex_fit_2_residuals	x	x	x	x
ve_ex_fit_residuals	x	x	x	
ve_ex_max	x	x	x	
ve_ex_min	x	x	x	
ve_ex_numTurns	x	x	x	
ve_ex_range	x	x	x	
ve_rmr_25th	x			
ve_rmr_50th	x			
ve_rmr_75th	x			
ve_rmr_avg	x			
ve_rmr_fit	x			
ve_rmr_fit_1	x			

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
ve_rmr_fit_1_residuals	x			
ve_rmr_fit_2	x			
ve_rmr_fit_2_residuals	x			
ve_rmr_fit_residuals	x			
ve_rmr_max	x			
ve_rmr_min	x			
ve_rmr_numTurns	x			
ve_rmr_range	x			
vo2_ex_25th	x	x	x	
vo2_ex_50th	x	x	x	
vo2_ex_75th	x	x	x	x
vo2_ex_avg	x	x	x	
vo2_ex_fit	x	x	x	x
vo2_ex_fit_1	x	x	x	x
vo2_ex_fit_1_residuals	x	x	x	x
vo2_ex_fit_2	x	x	x	x
vo2_ex_fit_2_residuals	x	x	x	x
vo2_ex_fit_residuals	x	x	x	x
vo2_ex_max	x	x	x	
vo2_ex_min	x	x	x	
vo2_ex_numTurns	x	x	x	x
vo2_ex_range	x	x	x	
vo2_rmr_25th	x			
vo2_rmr_50th	x			

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
vo2_rmr_75th	x			
vo2_rmr_avg	x			
vo2_rmr_fit	x			
vo2_rmr_fit_1	x			
vo2_rmr_fit_1_residuals	x	x		
vo2_rmr_fit_2	x	x		
vo2_rmr_fit_2_residuals	x	x		
vo2_rmr_fit_residuals	x			
vo2_rmr_max	x			
vo2_rmr_min	x			
vo2_rmr_numTurns	x	x	x	x
vo2_rmr_range	x	x	x	x
vo2kg_ex_25th	x	x	x	
vo2kg_ex_50th	x	x	x	
vo2kg_ex_75th	x	x	x	
vo2kg_ex_avg	x	x	x	
vo2kg_ex_fit	x	x	x	
vo2kg_ex_fit_1	x	x	x	
vo2kg_ex_fit_1_residuals	x	x		
vo2kg_ex_fit_2	x	x		
vo2kg_ex_fit_2_residuals	x	x		
vo2kg_ex_fit_residuals	x	x		
vo2kg_ex_max	x	x	x	
vo2kg_ex_min	x	x	x	

Table A.3: List of features included in each dataset

Feature	280 Features	197 Features	137 Features	67 Features
vo2kg_ex_numTurns	x	x	x	
vo2kg_ex_range	x	x	x	
vo2kg_rmr_25th	x			
vo2kg_rmr_50th	x			
vo2kg_rmr_75th	x	x	x	x
vo2kg_rmr_avg	x			
vo2kg_rmr_fit	x			
vo2kg_rmr_fit_1	x			
vo2kg_rmr_fit_1_residuals	x			
vo2kg_rmr_fit_2	x			
vo2kg_rmr_fit_2_residuals	x			
vo2kg_rmr_fit_residuals	x			
vo2kg_rmr_max	x			
vo2kg_rmr_min	x			
vo2kg_rmr_numTurns	x	x	x	x
vo2kg_rmr_range	x	x	x	x

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Centers for disease control and prevention - county data, 2013. URL `http://www.cdc.gov/diabetes/atlas/countydata/atlas.html`. [Online; accessed August 21, 2017].

[2] Centers for disease control and prevention, 2017. URL `http://www.cdc.gov/obesity/`. [Online; accessed August 21, 2017].

[3] A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz. The disease burden associated with overweight and obesity. *JAMA*, 282(16): 1523–1529, 1999.

[4] A. H. Mokdad, E. S. Ford, B. A. Bowman, W. H. Dietz, F. Vinicor, V. S. Bales, and J. S. Marks. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA*, 289(1):76–79, 2003.

[5] J. Cawley and C. Meyerhoefer. The medical care costs of obesity: an instrumental variables approach. *Journal of Health Economics*, 31(1):219–230, 2012.

[6] E. A. Finkelstein, J. G. Trogdon, J. W. Cohen, and W. Dietz. Annual medical spending attributable to obesity: payer-and service-specific estimates. *Health Affairs*, 28(5):w822–w831, 2009.

[7] C. N. Ochner, D. M. Barrios, C. D. Lee, and F. X. Pi-Sunyer. Biological mechanisms that promote weight regain following weight loss in obese humans. *Physiology & behavior*, 120:106–113, 2013.

[8] F. M. Sacks, G. A. Bray, V. J. Carey, S. R. Smith, D. H. Ryan, S. D. Anton, K. McManus, C. M. Champagne, L. M. Bishop, N. Laranjo, et al. Comparison of weight-loss diets with different compositions of fat, protein, and carbohydrates. *New England Journal of Medicine*, 360(9):859–873, 2009.

[9] K. Johansson, M. Neovius, and E. Hemmingsson. Effects of anti-obesity drugs, diet, and exercise on weight-loss maintenance after a very-low-calorie diet or low-calorie diet: a systematic review and meta-analysis of randomized controlled trials. *The American Journal of Clinical Nutrition*, 99(1):14–23, 2014.

[10] J. Pan and W. J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985.

[11] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):156–167, 2006.

[12] A. R. Relente and L. G. Sison. Characterization and adaptive filtering of motion artifacts in pulse oximetry using accelerometers. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, volume 2, pages 1769–1770. IEEE, 2002.

[13] A. E. Meuret, D. Rosenfield, F. H. Wilhelm, E. Zhou, A. Conrad, T. Ritz, and W. T. Roth. Do unexpected panic attacks occur spontaneously? *Biological Psychiatry*, 70(10):985–991, 2011.

[14] R. N. Bracewell and R. N. Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[15] K. I. Minami, H. Nakajima, and T. Toyoshima. Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network. *IEEE Transactions on Biomedical Engineering*, 46(2):179–185, 1999.

[16] W. S. Johnston and Y. Mendelson. Extracting breathing rate information from a wearable reflectance pulse oximeter sensor. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 2, pages 5388–5391. IEEE, 2004.

[17] K. Polat and S. Güneş. Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Applied Mathematics and Computation*, 187(2):1017–1026, 2007.

[18] B. Rigas, S. Morgello, I. S. Goldman, and P. Wong. Human colorectal cancers display abnormal Fourier-transform infrared spectra. *Proceedings of the National Academy of Sciences*, 87(20):8140–8144, 1990.

[19] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2):205–220, 1992.

[20] Y. Xu, J. B. Weaver, D. M. Healy, and J. Lu. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Transactions on Image Processing*, 3(6):747–758, 1994.

[21] V. Pichot, J. M. Gaspoz, S. Molliex, A. Antoniadis, T. Busso, F. Roche, F. Costes, L. Quintin, J. R. Lacour, and J. C. Barthélémy. Wavelet transform to quantify heart rate variability and to assess its instantaneous changes. *Journal of Applied Physiology*, 86(3):1081–1091, 1999.

[22] E. Martin. Novel method for stride length estimation with body area network accelerometers. In *Biomedical Wireless Technologies, Networks, and Sensing Systems (BioWireleSS), 2011 IEEE Topical Conference on*, pages 79–82. IEEE, 2011.

[23] A. N. Akansu, W. A. Serdijn, and I. W. Selesnick. Emerging applications of wavelets: A review. *Physical Communication*, 3(1):1–18, 2010.

[24] M. Kubat, R. C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3):195–215, 1998.

[25] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637, 2006.

[26] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[27] M. Dettling. BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593, 2004.

[28] Y. Bengio et al. Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1):1–127, 2009.

[29] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pages 342–350, 2009.

[30] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 265–272, 2011.

[31] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

[32] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003.

[33] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67: 93–104, 2012.

[34] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(1):319, 2008.

[35] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10(1):213, 2009.

[36] H. Wang and D. Hu. Comparison of SVM and LS-SVM for regression. In *Neural Networks and Brain, 2005. ICNN&B'05. International Conference on*, volume 1, pages 279–283. IEEE, 2005.

[37] T. Joachims. Making large-scale SVM learning practical. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.

[38] G. F. Smits and E. M. Jordaan. Improved SVM regression using mixtures of kernels. In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, volume 3, pages 2785–2790. IEEE, 2002.

[39] C. W. Hsu and C. J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[40] V. Guruswami and A. Sahai. Multiclass learning, boosting, and error-correcting codes. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 145–155. ACM, 1999.

[41] J. Zhu, H. Zou, S. Rosset, T. Hastie, et al. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[42] X. Li, L. Wang, and E. Sung. A study of AdaBoost with SVM based weak learners. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 1, pages 196–201. IEEE, 2005.

[43] W. Jiang. Process consistency for adaboost. *Annals of Statistics*, pages 13–29, 2004.

[44] A. Davies and Z. Ghahramani. The random forest kernel and other kernels for big data from random partitions. *arXiv preprint arXiv:1402.4293*, 2014.

[45] C. Vens and F. Costa. Random forest based feature induction. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 744–753. IEEE, 2011.

[46] G. Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[47] C. A. Biwer, A. E. Rothberg, H. IglayReger, C. F. Burant, and K. Najarian. Predicting medication regimen reduction in overweight and obese individuals. *BMC Obesity*, Submitted.

[48] C. Biwer, A. Rothberg, H. IglayReger, H. Derksen, C. F. Burant, and K. Najarian. Windowed persistent homology: A topological signal processing algorithm applied to clinical obesity data. *PLOS ONE*, 12(5):e0177696, 2017.

[49] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[50] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[51] H. Edelsbrunner and J. Harer. Persistent homology-a survey. *Contemporary Mathematics*, 453:257–282, 2008.

[52] J. E. Goodman, J. Pach, and R. Pollack. *Surveys on discrete and computational geometry: twenty years later: AMS-IMS-SIAM Joint Summer Research Conference, June 18-22, 2006, Snowbird, Utah*, volume 453. American Mathematical Soc., 2008.

[53] J. Roe. What is a coarse space. *Notices of the AMS*, 53(6):668–669, 2006.

[54] M. Guillemard and A. Iske. On groupoid C*-algebras, persistent homology and time-frequency analysis. *preprint*, 105, 2011.

[55] M. P. Dubuisson and A. K. Jain. A modified Hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568. IEEE, 1994.

[56] S. Emrani, T. Gentimis, and H. Krim. Persistent homology of delay embeddings and its application to wheeze detection. *IEEE Signal Processing Letters*, 21(4): 459–463, 2014.

[57] J. A. Perea, A. Deckard, S. B. Haase, and J. Harer. Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC Bioinformatics*, 16(1):257, 2015.

[58] F. A. Khasawneh and E. Munch. Stability determination in turning using persistent homology and time series analysis. In *ASME 2014 International Mechanical Engineering Congress and Exposition*, pages V04BT04A038–V04BT04A038. American Society of Mechanical Engineers, 2014.

[59] S. Emrani, T. S. Saponas, D. Morris, and H. Krim. A novel framework for pulse pressure wave analysis using persistent homology. *IEEE Signal Processing Letters*, 22(11):1879–1883, 2015.

[60] C. M. M. Pereira and R. F. de Mello. Persistent homology for time series and spatial data clustering. *Expert Systems with Applications*, 42(15):6026–6038, 2015.

[61] F. A. Khasawneh and E. Munch. Chatter detection in turning using persistent homology. *Mechanical Systems and Signal Processing*, 70:527–541, 2016.

[62] P. Bendich, E. Gasparovic, C. J. Tralie, and J. Harer. Scaffoldings and spines: Organizing high-dimensional data using cover trees, local principal component analysis, and persistent homology. *arXiv preprint arXiv:1602.06245*, 2016.

[63] K. Turner, S. Mukherjee, and D. M. Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344, 2014.

[64] R. Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[65] M. Kerber, D. Morozov, and A. Nigmetov. Geometry helps to compare persistence diagrams. In *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 103–112. SIAM, 2016.

[66] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, et al. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6): 2301–2339, 2014.

[67] K. Englehart, B. Hudgins, P. A. Parker, and M. Stevenson. Classification of the myoelectric signal using time-frequency based representations. *Medical Engineering & Physics*, 21(6):431–438, 1999.

[68] M. I. Ibrahimy, F. Ahmed, M. M. Ali, and E. Zahedi. Real-time signal processing for fetal heart rate monitoring. *IEEE Transactions on Biomedical Engineering*, 50(2):258–261, 2003.

[69] S. Q. Shi, T. S. Ansari, O. P. McGuinness, D. H. Wasserman, and C. H. Johnson. Circadian disruption leads to insulin resistance and obesity. *Current Biology*, 23 (5):372–381, 2013.

[70] J. Bass and J. S. Takahashi. Circadian integration of metabolism and energetics. *Science*, 330(6009):1349–1354, 2010.

[71] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[72] I. Rish. An empirical study of the naïve Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3(22):41–46, 2001.

[73] Diabetes Prevention Program Research Group et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine*, 2002(346):393–403, 2002.

[74] Look AHEAD Research Group et al. The Look AHEAD study: a description of the lifestyle intervention and the evidence supporting it. *Obesity (Silver Spring, Md.)*, 14(5):737, 2006.

[75] C. A. Biwer, N. Sohaee, A. E. Rothberg, H. B. IglayReger, H. Derksen, C. F. Burant, and K. Najarian. A graph-based approach to predicting the effects of weight loss on prescribed medications. *IEEE Journal of Biomedical and Health Informatics*, Submitted.

[76] N. Rohland and D. Reich. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, 22(5):939–946, 2012.

[77] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187, 2011.

[78] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.

[79] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[80] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245–271, 1997.

[81] S. B. Kotsiantis. Supervised machine learning: a review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real World AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24. IOS Press, 2007.

[82] E. W. Weisstein. Laplacian matrix, 1999. URL `http://mathworld.wolfram.com/LaplacianMatrix.html`. [Online; accessed August 21, 2017].