Neuromuscular Clinical Decision Support using Motor Unit Potentials Characterized by 'Pattern Discovery'

by

Lou Joseph Pino

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2008

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Objectives: Based on the analysis of electromyographic (EMG) data muscles are often characterized as normal or affected by a neuromuscular disease process. A clinical decision support system (CDSS) for the electrophysiological characterization of muscles by analyzing motor unit potentials (MUPs) was developed to assist physicians and researchers with the diagnosis, treatment & management of neuromuscular disorders and analyzed against criteria for use in a clinical setting.

Methods: Quantitative MUP data extracted from various muscles from control subjects and patients from a number of clinics was used to compare the sensitivity, specificity, and accuracy of a number of different clinical decision support methods. The CDSS developed in this work known as AMC-PD has three components: MUP characterization using Pattern Discovery (PD), muscle characterization by taking the average of MUP characterizations and calibrated muscle characterizations.

Results: The results demonstrated that AMC-PD achieved higher accuracy than conventional means and outlier analysis. Duration, thickness and number of turns were the most discriminative MUP features for characterizing the muscles studied in this work.

Conclusions: AMC-PD achieved higher accuracy than conventional means and outlier analysis. Muscle characterization performed using AMC-PD can facilitate the determination of "possible", "probable", or "definite" levels of disease whereas the conventional means and outlier methods can only provide a dichotomous "normal" or "abnormal" decision. Therefore, AMC-PD can be directly used to support clinical decisions related to initial diagnosis as well as treatment and management over time. Decisions are based on facts and not impressions giving electromyography a more reliable role in the diagnosis, management, and treatment of neuromuscular disorders. AMC-PD based calibrated muscle characterization can help make electrophysiological examinations more accurate and objective.

Acknowledgements

I would like to thank Daniel Stashuk for many fruitful discussions, thorough review of my results and manuscripts, and especially for providing sound advice. Dan works hard to help his graduate students succeed. Shaun Boe and Tim Doherty deserve praise for helping me understand the clinical environment where this work will be applied. Tim Doherty 'initiated' me to electromyography when I volunteered to be his subject for intra-muscular needle examination.

I would also like to thank Simon Podnar for sharing all of his MUP data, knowledge and collaboration towards an accepted manuscript. I would also like to thank Miki Nikolic for sharing his EMG data as well.

I would also like to thank my wife, Nancy, for understanding while I embarked on this career change.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

A clinical decision support method for the electrophysiological characterization of muscles to assist physicians and researchers with the diagnosis, treatment & management of neuromuscular disorders is developed and analyzed against criteria for use in a clinical setting. The method is transparent, and more accurate than the conventional decision support discussed in the literature.

1.1 Electromyographic Examination

Muscles are composed of groups of motor units. A motor unit (MU) is a motor neuron and the muscle fibers it innervates. A motor neuron is composed of dendrites that attach to other neurons, a cell body that houses a single nucleus and a single axon that terminates into multiple axonal twigs - each twig synapses with a muscle fiber in a region known as the neuromuscular junction (NMJ).

A motor neuron propagates impulses from the brain or spinal cord to muscle fibers to facilitate contraction of a MU. The top portion of Figure 1.1 shows a representation of the anatomy of a motor unit and a needle electrode in close proximity to the motor unit. The depolarization of a motor neuron will depolarize all of the muscle fibers that it innervates. The depolarization of a muscle fiber is known as a muscle fiber potential (MFP). The sum of all of the spatially and temporally dispersed MFPs arising from a depolarized MU generates changes in the voltage field in the extracellular volume surrounding the MU. The size and shape of the voltage field detected over time is a function of the morphology, physiology and the position of the electrode relative to the MU. The voltage field detected from an active MU is known as a motor unit potential (MUP). A series of depolarizations generated by one motor neuron (shown as Dirac Delta Impulse Trains in Figure 1.1) will generate a series of MUPs which is known as a motor unit potential train (MUPT). The superposition of voltage fields generated by active MUs that are detected by a needle or surface electrode is called the

Physiological electromyographic (EMG) signal in Figure 1.1. Figure 1.1 shows that the number of MUPTs detected is p. The Physiological EMG signal is denoted by $m_p(t, F)$ which is a function of time (t) and force of muscular contraction (F). Instrumentation introduces noise denoted by $n(t)$ and the detected EMG signal is also affected by the filter characteristics denoted by $r(t)$ of the instrumentation which leads to the observable EMG signal denoted $m(t, F)$ at the bottom of Figure 1.1. Neuromuscular disorders change the morphology and physiology of MUs causing changes in their activation patterns and MUP shapes and thus the EMG signals that they produce.

There are two broad categories of disorders that affect neuromuscular systems: myopathic and neuropathic. Myopathic disorders occur when muscle fibers die or atrophy, e.g. Muscular Dystrophy. Neuropathic disorders occur when motor neurons die and the remaining surviving motor neurons re-innervate orphaned muscle fibers, e.g. Amyotrophic Lateral Sclerosis (Lou Gehrig's disease). Figure 1.2 shows that in general the size and shape of MUPs detected from muscles affected by a myopathic disorder are smaller and more complex while the shapes of MUPs detected from neuropathic muscles are larger and more complex compared to MUPs detected from normal muscles. However, in practice there is a great deal of ambiguity in the interpretation of MUPs detected from a muscle as MUPs that appear normal, myopathic and neuropathic can be detected from any muscle regardless of its condition [1]. This ambiguity exists because of how widely muscle structure and needle position can vary.

In current clinical practice the status quo characterization of muscle is mostly done using qualitative auditory and visual analysis of needle-detected EMG signals detected during low-level muscle contractions and focuses on the analysis of individual MUP shapes and MU discharge patterns. In auditory analysis a clinician listens for the frequency and amplitude of the clicks and crackles made by amplified EMG signals. The auditory patterns reflect the recruitment of MUs and

the shape of MUP waveforms - both of which can be used to infer the presence or absence of underlying disease processes. Auditory analysis is usually done alongside a visual inspection of the EMG signal. Visual inspection tools use triggering in an attempt to isolate MUPs detected from the same MU. The triggered MUP waveforms are superimposed and displayed in real time. Qualitative visual inspection is a rough eyeball estimate of the shape of MUPs and how these compare to the expectation of being normal. According to several psychological studies qualitative examination is ambiguous and prone to misinterpretation [2, 3]. Both of these qualitative methods are crude with poor sensitivity and specificity, as well, they are lacking in providing objective quantitative data from which longitudinal comparisons can be easily made. For example, a recent study authored by Kendall showed that faculty and residents (blind to the underlying diagnosis of radiculopathy) using video recorded needle based examinations had an overall agreement of only 46.9% with the actual diagnosis [4].

Quantitative electromyography (QEMG) is the process of detecting and quantitatively analyzing EMG signals for the extraction of clinically useful information. Through the use of EMG signal decomposition a comprehensive set of features can be accurately measured and displayed leaving a large amount of information to be interpreted. QEMG is not used as frequently as qualitative EMG in clinics. The author believes that interpretation of the exhaustive set of statistics generated by quantitative analysis of EMG signals can be improved by transforming the statistics into clinically useful knowledge.

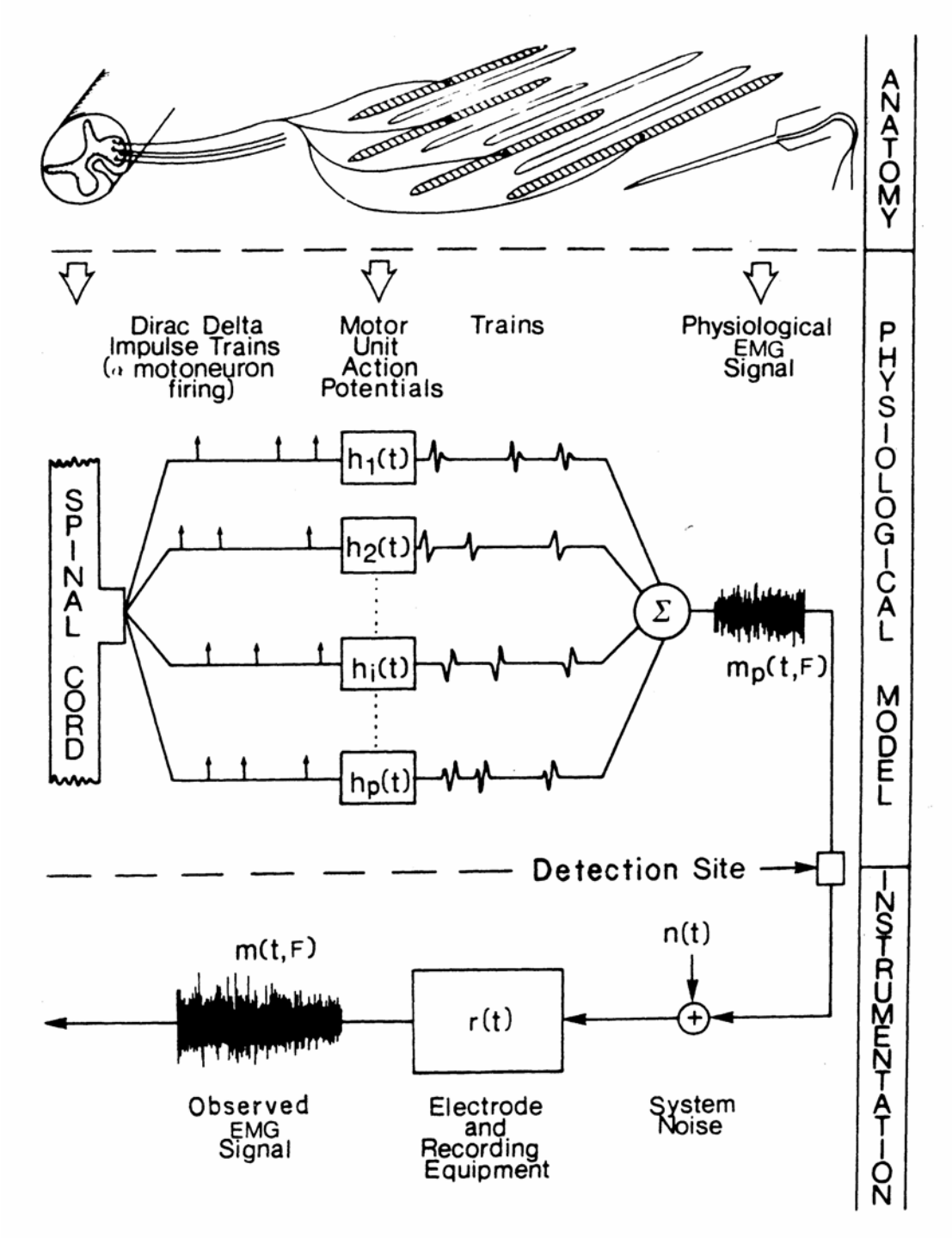


Figure 1.1 Model of the composition of an EMG signal. (From J. V. Basmajian, Muscles Alive: Their Functions Revealed by Electromyography. ,4 th ed. - ed. Baltimore: Williams & Wilkins, 1978, pp. xi, 495 p.)

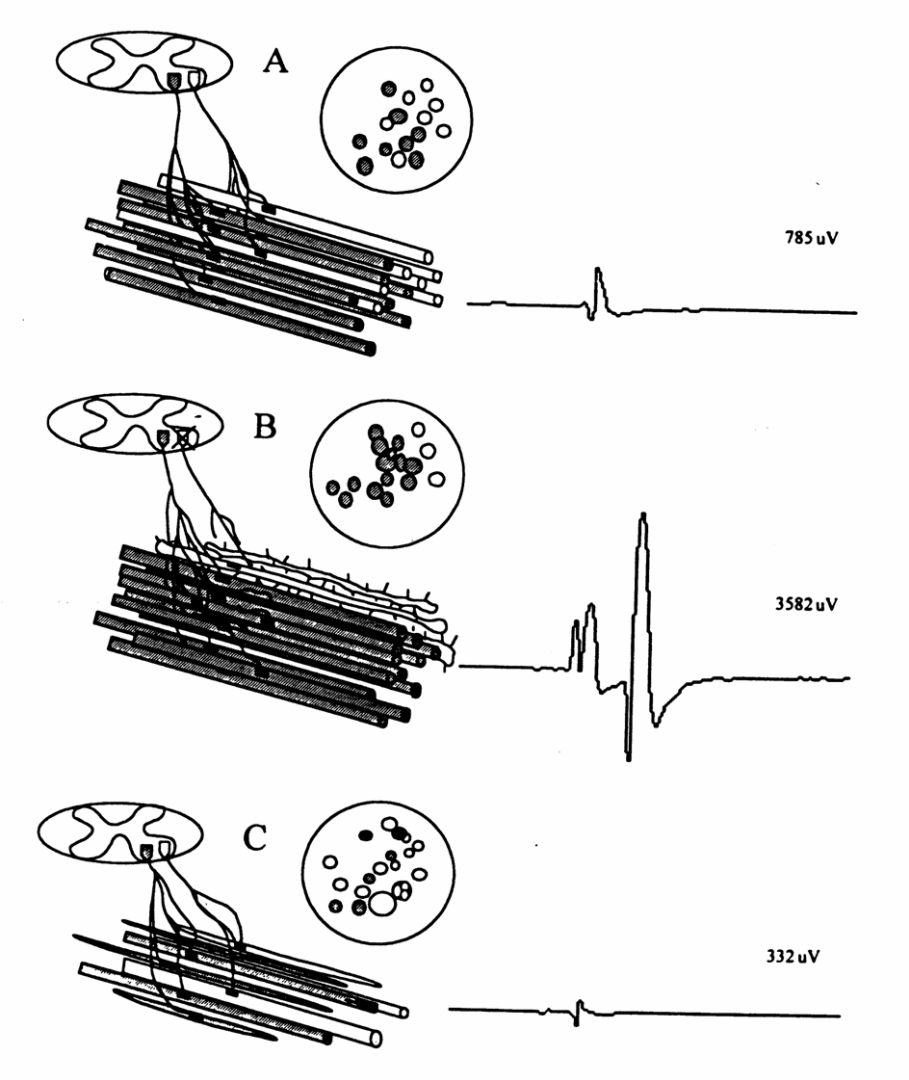


Figure 1.2 Effects of Disease on Motor Units & Motor Unit Potentials

Below the letter *A* in the figure is an example of two normal MUs and a MUP detected from one of the MUs. The circle to the right of the letter *A* is the cross-section of the muscle showing the individual muscle fibres belonging to each MU. Below the letter *B* is an example of two MUs being affected by a neuropathic disorder and the MUP detected from the MU shaded dark. Below the letter *C* is an example of two MUs being affected by a myopathic disorder and the MUP detected from the MU shaded dark. (From E. Stalberg and B. Falck, "The role of electromyography in neurology," Electroencephalography and Clinical Neurophysiology, vol. 103, pp. 579-598, 12. 1997.)

1.2 Objective & Approach

Characterization is a clinical term referring to the discernment, description or attribution of distinguishing traits [5]. Figure 1.3 below shows the information flow in a QEMG based examination that is to be augmented by the proposed Clinical Decision Support System (CDSS). Starting at the top a disease process may affect a muscle's morphology and physiology. An EMG signal is detected using an electrode that is inserted into a muscle. The signal is detected while the patient voluntarily contracts the muscle being examined. The method of QEMG analysis to be used in this work is the process of isolating MUPs that comprise the EMG signal and then extracting features of the shapes of the isolated MUPs. The step after QEMG shows the CDSS that provides a characterization of the extracted features for the muscle from which they were detected. It is shaded to show that this is the focus of the proposed research. The physician analyzes the characterization and then can infer if a disease is affecting a muscle and its level of involvement. The effect of a disease over time will be able to be tracked with the aid of the proposed CDSS.

The objective of this work is to augment existing QEMG techniques with a decision support system that transforms QEMG generated statistics into a concise muscle characterization that improves the ability of physicians to decide upon an appropriate clinical action. Pattern recognition techniques will be used to develop a characterization system. A system that has excellent sensitivity (true positives) and specificity (true negatives) when used to characterize muscle disorders would be a valuable addition to clinical practice.

The CDSS developed in this work analyzes QEMG statistics to report the characterization of a single muscle. The CDSS has three components: MUP characterization, muscle characterization and calibrated muscle characterization. MUP characterization calculates a set of conditional probabilities that a MUP is abnormal and normal given the features of the MUP detected from a muscle under

examination. Combining a set of MUP characterizations provides an overall muscle characterization. A muscle characterization is a set of scores representing the degree to which a muscle is normal and abnormal. The third component of the CDSS converts scores into calibrated muscle characterization conditional probabilities that accurately reflect the reliability or confidence of an individual characterization. In other words, a calibrated muscle characterization score of 80% probability of abnormality means that 80% of muscles that receive that conditional probability are truly abnormal. Calibrated conditional probabilities will help clinicians understand the level of confidence a CDSS has in its categorization of normality and abnormality of a muscle under examination.

1.3 Overview of the Thesis

Chapter 2 reviews QEMG methods used for muscle examination in clinical settings. The different methods are compared to justify focusing on MUP analysis. Then existing MUP characterization methods are described as well as existing muscle characterization methods that use MUP features. A section in Chapter 2 provides the requirements that CDSSs in general need to possess. These requirements guided the development and evaluation of the CDSS in this work. Chapter 2 ends by describing potential pattern recognition techniques and discusses their suitably for use as MUP characterization methods against the CDSS requirements.

Chapter 3 describes the various MUP data sets that were used to evaluate the CDSS developed for this work. The set of MUP features are defined as well as the measures used to evaluate the performance of the CDSS.

Chapter 4 describes the MUP Characterization method used in this work known as Pattern Discovery (PD). The performance of PD was compared to other classifiers that are commonly used for medical diagnosis.

Chapter 5 analyzes the performance of various methods used for muscle characterization. Combining conditional probabilities, calculated by pattern recognition based MUP Characterization methods, into muscle characterizations are known as probabilistic muscle characterization methods. The performance of a number of probabilistic muscle characterization methods are compared with conventional muscle characterization techniques across an exhaustive set of feature sets and for the various MUP data sets.

Chapter 6 describes another muscle characterization method – the Z transform. Then a method of calibrating muscle characterizations that converts raw muscle characterization scores into conditional probabilities is described. The performance of the calibration method was compared across two different muscle characterization methods, AMC-PD and ZT-PD using all of the MUP data sets.

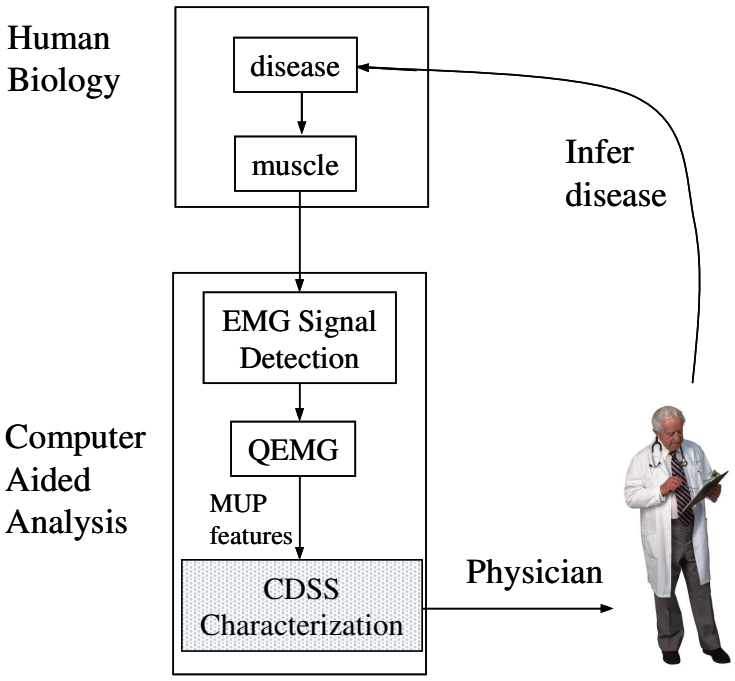


Figure 1.3 Information Flow to Characterize a Muscle

Chapter 2
Background and Related Work

2.1 Quantitative EMG Overview

Quantitative EMG (QEMG) is the use of quantitative based methods to analyze and extract features from EMG signals to augment the determination of underlying disease processes. QEMG analysis offers objective, quantitative analysis that can allow comparisons with reference data collected from subjects of the same age and gender as the patient. There are four main approaches for quantitatively studying the electrical activity of muscles contracted voluntarily by the patient. The first approach known as MUP analysis is based on studying the shapes and sizes of isolated MUPs. The second approach is based on studying the composite EMG signal also known as the interference pattern (IP). The third approach estimates the number of motor units in a muscle. The fourth examines the stability of the operation of the neuromuscular junction.

2.1.1 MUP Analysis

The contraction of a muscle leads to the discharge of recruited motor units. A sustained contraction at a constant level of force leads to the repeated discharge of these MUs. The series of MUPs associated with a MU is known as a MUP train. The composite EMG signal is the superposition of all of the MUP trains of active MUs. Neuromuscular disorders change the morphology and physiology of MUs causing changes in their activation patterns and MUP shape and size. As well the orientation and proximity of the electrode relative to the active MUs affects the size and shape of detected MUPs.

Several techniques are available for the extraction and analysis of isolated MUP waveforms. One that is in current widespread use is amplitude triggering. MUPs that exceed an amplitude threshold can be triggered for view on a screen. This method is used to capture a set of MUPs that exceeds an amplitude threshold. To capture MUPs of differing amplitude a window trigger mechanism was

developed [6]. MUPs whose peak amplitude falls within a specified range can be isolated for view on a screen. Repeated triggering of MUPs that fall within this range can be averaged to remove noise from interfering MUPs not belonging to the MUP train of interest. However, MUPs detected from different MUs whose peak amplitude falls in the same range may also trigger the display making it harder to isolate MUP trains. Amplitude triggering requires a great deal of time and patient cooperation if a large number of MUPs are to be sampled.

Another method of isolating MUPs is known as decomposition based MUP analysis. This method involves the detection, clustering and then the supervised classification of MUPs into MUP trains [7]. A MUP is detected when it exceeds threshold levels - usually based on amplitude and slope. MUPs in the first several seconds (initialization interval) of the EMG signal are clustered using the K-means algorithm. The clustering algorithm produces an estimate for the number of active motor units and a prototypical shape of their MUPs. Supervised classification places each detected MUP into a MUP train based on a certainty measure. In addition to using shape information the expected firing behaviour of MUs are estimated and used to determine if a candidate MUP belongs to a MUP train. In this way MUPs with similar shapes but created by different MUs can be differentiated by comparing their respective patterns of occurrence. MUPs classified as belonging to the same train that exceed a certainty threshold are averaged to form the MUP template for that train. It is this template that is measured to extract features representative of that MUP train.

2.1.2 Interference Pattern Assessment

EMG signals detected from a muscle during moderate to high levels of contraction are composed of large numbers of MUPTs that are superimposed. As such, an EMG signal is also known as an Interference Pattern (IP). The density and amplitude of an IP can be used to estimate MU recruitment

and firing rates. There are two computer-aided methods for analyzing an IP: frequency and time domain analysis [8].

Frequency Domain Analysis

The power spectrum of an IP can be used to infer features from the MUPs that comprise the IP. MUPs detected from myopathic MUs tend to be complex, have short durations and fast rise times as compared to MUPs indicative of normal MUPs. This will be reflected in the power spectrum as higher frequency components. MUPs detected from neuropathic MUs tend to have long durations and slow rise times as compared to MUPs indicative of normal MUPs. This will be reflected in the power spectrum as lower frequency components. The specificity and sensitivity of frequency domain analysis is poorer than time domain analysis because of the higher variability of an IP's frequency spectrum.

Time Domain Analysis

Time domain analysis is based on the number of local peaks and valleys and the amplitude between successive peaks of opposite polarity. A peak is a turn if the amplitude to the next peak of opposite polarity exceeds a threshold of 25 or 50 μV. The amplitude of a turn is the difference in voltage between successive peaks of opposite polarity. One method of turns analysis considers the ratio of the number of turns (NT) to the mean amplitude (MA) of all of the turns. In general, a high NT/MA ratio is indicative of myopathy and a low ratio is indicative of neuropathy. Figure 2.1 provides a diagram that helps to explain how MA and NT are calculated.

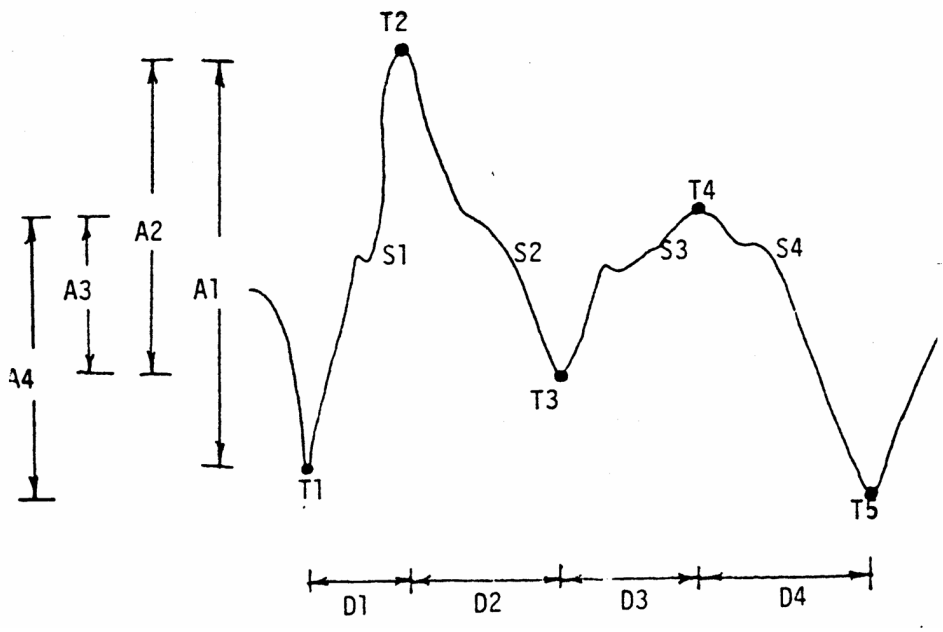


Figure 2.1 Time Domain Analysis of an IP Signal

A letter *T* with a number refers to a turn. A letter *S* with a number refers to a segment of the IP that occurs between two successive turns. A letter *A* with a number refers to the amplitude of a segment. So as an example, A1 is the amplitude of S1, i.e. the segment from turn 1 (T1) to turn 2 (T2). Mean amplitude (MA) is calculated across all segments and NT is the total number of turns. (From SD Nandedkar, DB Sanders, EV Stalberg: Simulation and analysis of the electromyographic interference pattern in normal muscle. Part II: Activity, upper centile amplitude, and number of small segments. Muscle Nerve. 1986 Jul-Aug;9(6):486-90).

2.1.3 Motor Unit Number Estimation

Neuropathic disorders can decrease the number of motor units in muscles. An estimate of the number of motor units in a muscle suspected of a neuropathic disorder can be useful in tracking the progress of a disorder and or effectiveness of its treatment. A pair of EMG signals, one detected using a surface electrode and the other using an intramuscular needle electrode, are used together to estimate

the number of motor units in a muscle. A motor nerve bundle is electrically stimulated with sufficient intensity to ensure all motor neurons leading to the muscle under test are depolarized. The surface EMG signal detected above the muscle under test is called a Compound Muscle Action Potential (CMAP). MUPs detected in the intramuscular needle EMG signal underneath the surface electrode are classified into separate trains where each train is believed to be detected from the same motor unit. The set of firings of intramuscular detected MUPs belonging to the same train are used as triggers for locating intervals in the surface-detected signal. The set of intervals are ensemble averaged to extract the surface MUP (SMUP) corresponding to the intramuscular MUP train. The peak-to-peak amplitude of each SMUP is calculated. The average peak-to-peak amplitude is determined across all of the SMUPs. The number of motor units in the muscle under test is estimated by dividing the peak-to-peak amplitude of the CMAP signal by the average peak-to-peak amplitude of the SMUPs [7].

2.1.4 Muscle Fiber Jitter

Jitter is measured by using the action potential of one muscle fiber as a reference and measuring the variability in time of the action potential of a second muscle fiber from the same motor unit [9]. Increased jitter is a clinical sign of a defective neuromuscular junction (NMJ). In a normal NMJ, there is an excess of acetylcholine (ACh) in the pre-synaptic terminal that can bind with numerous healthy post-synaptic ACh receptors leading to a depolarization of the muscle fiber. A defective NMJ occurs when either ACh or healthy ACh receptors are lacking leading to increased variability in time required to bind sufficient ACh to the post-synaptic membrane in sufficient quantities to depolarize the muscle fiber. In some cases, the number of ACh bindings is insufficient to cause a depolarization to occur at all thus blocking is said to occur.

Increased jitter is often a sign of myasthenia gravis - a disorder that affects NMJs. Myasthenia gravis is caused by circulating antibodies that inhibit ACh receptors in the post-synaptic membrane from binding with ACh.

2.1.5 Comparison of the QEMG Techniques

Isolated MUP analysis can provide direct information about MU size, MU fiber density, the performance of neuromuscular junctions, the temporal dispersion of muscle fiber action potentials and MU firing behaviour. Time domain based IP EMG analysis can provide indirect information about the size of MUs. IP QEMG can be useful for examining disease processes that affect MUs that are active only at higher levels of force since isolated MUP analysis becomes more difficult at higher levels of contraction. However for most disease processes isolated MUP QEMG provides a more comprehensive set of MU features [8].

Motor Unit Number Estimation (MUNE) is useful when a neuropathic disorder is known or suspected to evaluate the extent of MU loss and to determine the effectiveness of treatments. MUNE is not likely to be used for an initial electrophysiological examination if the state of a patient's neuromuscular systems is unknown, i.e. normal, myopathic or neuropathic or if the specific muscles affected in a patient are unknown.

Analysis of jitter is only useful for diseases affecting the NMJ.

2.2 MUP and Neuromuscular Characterization

Section 2.1 gave an overview of QEMG techniques. This section provides further background for the MUP analysis method (2.1.1) used as the basis for the development and analysis of the CDSS which is the focus of this thesis. This work will use statistics generated by decomposition based MUP analysis, specifically features associated with MUP size and shape and use these feature values for the

characterization of MUPs. Other QEMG methods can be included in future work as sources of additional information for determining an overall muscle characterization.

In general, QEMG decomposition based examination can be thought of as having three phases: 1) Data Acquisition, 2) Information Extraction: measurements and statistics determined from the data and 3) Interpretation: interpreting the information to reach conclusions. The interpretation phase can be thought of as having three steps: a) characterization of individual MUPs; b) integration of individual MUP characterizations into a muscle characterization and c) inference by a clinician as to whether a disorder is affecting the muscle or not, and if so, the type of disorder (myopathic or neuropathic) and its level of involvement. In contrast to the qualitative interpretation currently used in QEMG examinations, quantitative interpretation can facilitate the objective measurement of the degree of involvement of a neuromuscular disorder. This work addresses steps 3 a) and b) of the interpretation phase by describing and evaluating characterization processes that provide quantitative interpretation of information commonly extracted from individual MUPs during a QEMG examination.

2.2.1 Existing MUP Characterization Techniques

This work uses automated interpretation of quantitative shape-based isolated MUP features for the characterization of a muscle. Methods in the literature for the characterization of individual MUPs, i.e. step 3 a) of the interpretation phase of a QEMG examination, are described in this section. Typically, the literature discusses how feature values for a set of MUPs detected from a single muscle can be combined into a muscle characterization using just three categories: myopathic, normal or neuropathic. These three categories provide an initial step towards a useful, robust neuromuscular CDSS.

The literature describes a number of different processes for characterizing MUPs. Pattichis et al. [10] applied artificial neural network (ANN) models to the classification of MUPs sampled from the

biceps brachii muscle of 14 normal patients, 16 neuropathic and 14 myopathic patients. The data was divided into 24 training sets and 20 test sets. For each set, the means and standard deviations of 7 features (duration, spike duration, amplitude, area, spike area, number of phases and number of turns) of the MUPs belonging to each muscle were determined. An error rate of 10 to 20% was achieved depending on the specific ANN architecture used. For instance, the ANN architecture that had 40 neurons in the first hidden layer and 10 neurons in the second hidden layer achieved an error rate of 10%. The authors found that ANNs easily tended towards over-fitting i.e., it was difficult to achieve generalization, the ability of the ANN to correctly classify unknown cases based on the training data. ANN models are not transparent because they do not reveal how they reach their conclusions. The large number of neurons means that a large number of arithmetic operations are used to transform the features making ANNs essentially black box classifiers.

Pfeiffer and Kunze [1, 11] used Fisher's Linear Discriminant Analysis (LDA) to determine the probability of the abnormality of each MUP (i.e. MUP characterization). The technique estimated the conditional probabilities (one for each category) of a MUP being detected from a muscle with a given category of disorder.

In [1, 11], the discriminant model was trained based on 363 MUPs from 15 normal muscles, 467 MUPs from 16 myopathic muscles and 463 MUPs from 23 neuropathic muscles using duration, area, turns count and center frequency as the features. The error rates using a single MUP for muscle characterization were 52.9% for myopathic disorders, 44.2% for neuropathic disorders and 35.2% for normal muscles with sample sizes of 1676, 714 and 2636 MUPs respectively. According to Pfeiffer & Kunze [1, 11], the high error rate could have been reduced if samples acquired during periods, 2 years apart were done in a more consistent manner.

The method developed by Pfeiffer and Kunze determines a numeric probability of a MUP being detected from a normal, myopathic and neuropathic muscle. This allows a set of MUPs sampled from

a muscle to be quantitatively turned into a neuromuscular characterization. Despite the sampling problems experienced during testing, LDA is a successful method for MUP characterization. However, because of the arithmetic transformation of the features, LDA is not as transparent as classifiers that provide logical relationships between the features.

2.2.2 Existing Muscle Characterization Techniques

This section summarizes methods found in the literature describing how information across a set of MUPs detected from a muscle under test can be combined into a characterization of that muscle, i.e. step 3 b) of the interpretation phase of a QEMG examination. There are two main methods for combining MUP statistics into a muscle characterization: means/outlier analysis and probabilistic methods. The means/outlier method is done by comparing the mean of a set of MUP feature values that are below or above normative limits and or by counting the number of outliers [12, 13]. Probabilistic methods were first introduced by Pfeiffer [1] who used Bayes' rule to characterize muscles by combining for each MUP detected their conditional probabilities of being detected in a healthy muscle or one with a specific disorder. Analysis of individual MUP feature values lacks sufficient information to accurately characterize a muscle so Bayesian aggregation provides a statistically robust method for combining the feature values of several MUPs acquired from a muscle.

Stewart et al. developed and evaluated a computer based system for acquiring MUPs, measuring their features and inferring a diagnosis (or characterization) [13] called the Means method. The mean and standard deviations of amplitude, area and thickness (area-to-amplitude ratio) values for sets of MUPs acquired from 68 normal subjects (18 - 62 years old) were calculated. The percentage of polyphasic and polyturn MUPs detected from these subjects was calculated as well. The normal range was defined as mean +/- 2 standard deviations. The method classifies a muscle as myopathic if one or more of the mean feature values falls below the normal range and classifies the muscle as neuropathic if one or more of the features falls above the normal range. The method was tested with MUPs

detected from 50 patients with known myopathic and 55 patients with known neuropathic disorders. The error rate for characterizing patients as being affected by myopathic muscles was 44% and the neuropathic error rate was 36.7%.

Stalberg et al. developed the Outlier method for characterization of muscles [12]. The goal of the method was to increase accuracy and reduce the number of MUPs that needed to be collected during a qualitative EMG examination. The method uses either extremely low or high feature values of MUPs, i.e. outliers to determine abnormality. The method uses the 5^{th} and 95^{th} percentiles of the third smallest and third highest set of values per feature collected from muscles to establish the low and high outlier threshold values. A muscle under test is declared myopathic if it has three or more low outliers of the same feature value or neuropathic if it has three or more high outliers of the same feature value.

2.3 Generic Requirements for Clinical Decision Support Systems (CDSS)

The following requirements are based on ideas developed by Kononenko [14] and Sprogar et al. [15] who describe a set of requirements needed for machine learning systems used in medical decision support. These requirements were used to guide the development and evaluation of the muscle CDSS described in this work.

1) Transparent: Characterizations need to be presented in a manner that allows a clinician to understand how a characterization was determined. This is especially important when faced with an unexpected characterization that contradicts a physician's initial expectation or intuition. A physician's knowledge of the basis of a characterization is critical to the confidence in its veracity. A system's ability to explain its characterizations or output is an important part of a physician's acceptance of the system [16]. According to Feng [17] transparency to users requires that a characterization system provide logical versus arithmetic expressions of the features. Operators such as "and", "or" and "if-then" are preferred to provide connections between feature values used to

explain classifications. These expressions are valuable since they provide meaningful explanations to human experts and are more easily evaluated.

2) Accurate: Characterizations need to maximize both specificity and sensitivity beyond what is typically achieved through routine subjective analysis of an EMG signal. This is essential if the approach is to be considered useful by clinicians.

3) Report Confidence: A confidence measure reports the degree to which a characterization suggested by a classifier is likely to match the 'true' underlying characterization. It is intended to help clinicians minimize errors. A confidence measure is meant to minimize the number of situations where a characterization system makes a correct suggestion but is ignored by a physician. A confidence measure can also help avoid situations where a physician accepts an incorrect suggestion made by a characterization system if he/she over-trusts the characterization.

4) Numeric Characterization Value: Characterizations need to be presented as numeric measures supporting or refuting each category under consideration, e.g. probabilities. The numeric scale can be ordinal but it is preferably that it has continuous values. A numeric MUP characterization will allow a method to combine individual MUP characterizations into an overall neuromuscular characterization.

5) Mixed Mode Multi-Variate Features and Interdependency: Analyzing a single feature of a MUP has poor discriminatory power because of the wide range of values for any one MUP feature and the substantial overlap in the distributions of MUP features reflective of normal versus a diseased neuromuscular system [18]. Discriminatory power can be increased when multiple feature values are considered simultaneously [19]. For instance, a MUP with high amplitude only provides evidence that it was detected from a neuropathic muscle if it also has long duration.

MUP characterization must be able to handle various mixed mode multi-variate data types that include numeric, Boolean, nominal and or ordinal. Numeric data can be continuous or discrete. Nominal data values are non-numeric, descriptive or use categorical labels. For instance the patient is

"normal" is an example of nominal data. Ordinal data values define a position or rank. For instance the patient "is weak" is an example of ordinal data. The system needs to handle any underlying joint probability distribution of the features used for characterization. A classifier needs to capture dependencies among the features. Clinical patterns that offer important clues to the type of a neuromuscular disorder often combine different data types. For instance, proximal muscles (nominal) affected by weakness (ordinal via MRC strength scale) with little or no wasting (ordinal) are often an indication of a myopathy [20] (nominal). Nominal data types provide a means of capturing more data relevant to the examination. In cases where there is a great deal of imprecision or subjectivity, a nominal description can provide useful additional information. For instance, the grading of strength on the MRC scale of 0 to 5 can be augmented with a description such as the smoothness (or lack thereof) of the contraction.

6) Generalization: A MUP characterization system needs to accurately classify novel patterns that have not appeared in the training data. Classifiers that have been 'tuned' to the training data (i.e. over-fitting) can achieve low error rates based on testing of the training data. However, they are unlikely to perform well for MUPs that have not been seen before or were not included in the training data [19].

7) Handle Missing Data. A MUP characterization system needs to be able to handle missing feature values both in the training data and in the data extracted from the muscle under consideration without adversely affecting the outcome of a characterization.

2.4 Possible Classification Methods

Methods to characterize a neuromuscular system and to measure the degree of involvement of a disorder require a MUP characterization process that is accurate, allows the basis of it decisions to be easily understood, produces a numeric value in support of or refutation of a characterization, and is able to achieve generalization. Existing MUP and muscle Characterization techniques do not

completely satisfy these important requirements. It is hypothesized, that a classifier using pattern recognition techniques can meet these requirements. This section discusses the suitability of various classification methods based on the requirements in Section 2. 3.

2.4.1 1st Order Logic

A rule in first order logic is a sentence that is composed of quantifiers, variables, constants and functions that are connected with logical operators. A knowledge domain can be built up by asserting a set of rules that are known to be true. A query can be asked of the knowledge domain and a first order logic system will return true, false or unknown for a given query. Measurements that are continuous random variables must be converted to logic symbols using quantization techniques. In general it is difficult for a classifier based on 1st order logic to calculate probabilities. Therefore these systems are not suitable for the CDSS under consideration.

KANDID [21] is an example of a decision support system that uses first order logic whose acronym stands for Knowledge-based Assistant for Neuromuscular DIsorder Diagnosis. It is used to plan and manage nerve conduction studies as well as diagnosis and reporting of neuromuscular disorders.

2.4.2 Bayesian Networks

A Bayesian network is a directed graph where each node represents a feature. Sets of arrows connect pairs of nodes where an arrow from node X_1 to node X_2 means that X_1 is a parent of X_2. Each node stores a table of conditional probabilities where the parent nodes are the given features of a conditional probability. The conditional probability table associated with a node provides the effect of the state of the parents on the probability outcome of the node. A Bayesian network can provide an estimate of the joint probability density function for a classification problem. The joint probability of a set of feature values occurring is given by the notation $P(X_1 = v_1, X_2 = v_2, \ldots, X_n = v_n)$ where

$X_i = v_i$ means that feature X_i has taken on the value v_i. A probability prediction for $P(v_1, v_2, \ldots, v_n)$ can be calculated as follows [22]:

$$P(v_1, v_2, \ldots, v_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i)) \qquad (2.1)$$

where $parents(X_i)$ represents the values of the features of parent nodes of node X_i

An example of a muscle disorder decision support system based on Bayesian Networks is called MUNIN [23]. MUNIN stands for MUscle & Nerve Inference Network. There are four layers in its Bayesian Network leading from the underlying disorders to the findings. At the top layer are the muscular disorders. The second layer is the impacts of the disorders on the physiology of a muscle. The third layer is the impact on the physiology of a muscle as well. Layers two and three are both required for physiological factors in the event that more than one disorder is present at one time. The last layer represents the findings as determined by clinical tests, instruments, and doctor examination or patient reports. The major limitation of MUNIN is that it requires human medical expertise in determining the conditional probability tables. Machine learning techniques can be applied to learn the optimal structure of a Bayesian network from training data, however, the problem is considered intractable because of the large space of possible network structures to be searched [22, 24].

Machine learning is more efficient for Naïve Bayesian Networks. A Naïve Bayesian (NB) classifier is built by assuming that all of the features are conditionally independent of each other. A NB classifier is rather simplified in that it doesn't reveal any interdependencies among the features. The NB classifier produces a score s_k for each category by taking the product of the conditional probabilities of each feature whose value falls within a pre-defined interval [19]. The scores are normalized resulting in a set of conditional probabilities that sum to 1. It is a discrete classifier since it handles nominal and discrete data types and requires continuous features to be quantized. A NB

classifier cannot uncover the underlying distribution of the data nor easily provide a transparent explanation. Therefore a naïve Bayesian network is unlikely to uncover the underlying distribution of the data as well as a full network. NB classifiers can be augmented so that they provide some transparency by reporting the individual contribution of each feature towards the probability that an observation belongs to a category [25].

2.4.3 Decision Trees

A decision tree (DT) is a structure of tests done on the feature values of a test instance that leads to a classification. Each node in a tree specifies a test to be made on a single feature. The first node of the tree is known as the root node. Each branch coming out of the node is labeled with the outcome of the test applied at the node. When a leaf of the tree is reached the value specified by the leaf, i.e. the category is returned as the result to a query made using feature values. It can handle discrete features, Boolean, nominal and ordinal. Continuous features need to be quantized. A DT can produce a conditional probability estimate of the category given a test instance. Specific high order interactions are difficult to uncover because a decision tree is built by considering one feature at a time. DTs can be augmented so that rules can be extracted by tracing the path taken through the tree. Since a single feature at a time is tested this does not readily reveal relationships among the features [26].

There are a number of algorithms that will build decision trees based on labeled training data. One of the most popular is Quinlan's [27] ID3 algorithm for constructing decision trees. The discriminatory power of a feature can be determined by calculating the entropy [19] of the various categories that match the feature value that the emerging branch of the node represents. Entropy provided by a node can be thought as the impurity across all of the branches emerging from a node where each branch represents a specific feature value – higher entropy corresponds to higher levels of impurity. In a two category system impurity in a branch is the ratio of the number of instances of

training data of the category with the greatest number of instances over all instances of data appearing in that branch. So if each branch had only instances of training data belonging to a single category then each branch would have zero impurities, i.e. zero entropy. Therefore a feature where each branch has no impurities would contribute high classification accuracy to a DT. The process for inducing a decision tree from training data starts by calculating the impurity for each feature. The feature with the lowest level of impurity is chosen to be the first test in the tree since it has the highest classification accuracy. This algorithm is applied iteratively to the other features until the leaves of the tree are reached.

The C4.5 algorithm is an extension to ID3 that adds abilities such as handling missing data, pruning of decision trees, rule derivation and the quantization of feature value ranges [19]. These improvements lead to decision trees that are smaller and often achieve higher classification accuracy.

Decision trees can provide a conditional probability estimate of a category given a set of feature values by counting the number of instances of a category that appear in a leaf node. In general, decision trees provide poor probability estimates for two reasons [28]. First, decision tree induction methods try to make leaves homogeneous so observed probabilities in leaves are biased towards zero or one. Second, the estimates have high variance because often the numbers of training examples in a leaf are small leading to observed frequencies that are not statistically reliable.

A decision tree does not give insight into the relationship among features [26]. Specific high order interactions are difficult to uncover because a decision tree is built by splitting one variable at a time.

Physicians typically agree that the first node in a decision tree is the most important question [14]. The first node in a tree induced from training data may not reflect the most important diagnostic question because a feature is chosen based on reducing node impurity. Also a pruned tree does not allow a physician to look at the contribution of all of the features in making a diagnosis.

2.4.4 Artificial Neural Networks

A popular tool in pattern recognition is the artificial neural network (ANN). Inspired by an understanding of the human brain an artificial neuron is composed of input links, input function, an activation function and an output [22]. A single layer of neurons can be arranged to provide a single linear decision boundary. Multiple layers of neurons can be arranged to provide multiple linear decision boundaries and can be used to represent a wide range of decision functions. ANNs are not totally transparent in that they cannot easily provide an explanation for their conclusions that are directly understood.

2.4.5 Fisher's Linear Discriminant Analysis

Fisher's Linear Discriminant Analysis (LDA) is an approach that selects a set of features associated to each category, that can maximally distinguish among different categories or classes i.e. maximize the ratio of (variance between categories) / (variance within categories) [29]. The coefficients of the functions are unit vectors that maximize the above ratio. The number of discriminant functions is: min (#classes-1, #parameters). So with *K* number of classes there are (*K*-1) number of discriminant functions, unless the number of features is less than (*K*-1). Discriminant functions rotate and scale the training data to maximize the distance between categories. Each instance in a set of training data can be transformed into a set of discriminant coordinates using the learned discriminant functions. The center of a category can be estimated by calculating the mean of the discriminate coordinates of a set of training data. The distance to the category centroid is used to estimate whether a test instance belongs to a specific category or not. An example of two discriminant functions called D_1 and D_2 with three features F_{1-3} are shown below.

$$D_1 = v_{11}*F_1 + v_{21}*F_2 + v_{31}*F_3 \quad (2.2)$$

$$D_2 = v_{12}*F_1 + v_{22}*F_2 + v_{32}*F_3 \quad (2.3)$$

v_{ij} is the coefficient for the i^{th} feature for the j^{th} discriminant function

The coefficients v_{ij} are determined by maximizing the covariance between the categories in the training data over the covariance within the categories. This is subject to the condition that the coefficients *v* are unit vectors otherwise the maximum is unbounded.

$$\frac{v^T B v}{v^T W v} = \lambda \qquad (2.4)$$

B is the between categories covariance matrix and ***W*** is the within categories covariance matrix. The ***B*** & ***W*** matrices can be calculated from the training data. Taking the first derivative of the above equation with respect to *v* and solving leads to the solution of *v*.

Pfeiffer and Kunze [1, 11] applied LDA to characterize MUPs by calculating the probability that a particular MUP was detected from a myopathic, normal, or neuropathic muscle. Each MUP in the training data is mapped to a set of discriminant coordinates (D_1, D_2) by applying the discriminant functions to the feature values of the MUPs. Test MUPs are classified by finding the Euclidean distance to the centroid of the set of discriminant coordinates in the training data of each category. The test MUP is classified to the category whose centroid has the minimum distance to the test MUP.

2.4.6 Pattern Discovery

Information-theory based statistical inference can be used for the detection of significant patterns. Wong and Wang [30-33] describe the use of information theory based pattern discovery (PD) for classification. Their classification algorithm is composed of three parts: discovery of significant patterns, rule selection and classification.

A significant pattern is said to occur when a set of feature values occurs together more often than expected assuming a random occurrence. For continuous data, a range of feature values is quantized into a natural language label as is used in the clinic: e.g. small, medium or large. A statistical test of significance determines if a significant pattern has occurred. Patterns that include a category label can be used as rules for classification. The order of a rule is determined by the number of features, including the category label, found in a pattern. An example of a 4^{th} order rule is a MUP 1:labeled myopathic with 2: low amplitude, 3: low duration and 4: high number of turns.

An information theoretic measure called weight of evidence (WOE) represents the discriminatory power of a rule. A rule may be statistically significant but the WOE is needed to determine if the rule provides negative, neutral or positive evidence to refute or support a classification. A large negative WOE score means that the category suggested by the rule is highly unlikely to occur. A large positive WOE score means that the category suggested by the rule is highly likely to occur. Rules with WOE scores near zero neither support nor refute the likelihood of the classification. Section 4.1 will discuss this technique in more detail.

The rules used for classifying an observation are selected starting with the highest order rule first and accumulating its WOE. The features used in the highest order rule are then excluded and the rule with an equal or lower order is found and its WOE is added to the total. This process is continued until no more rules are found or all of the features have been considered. PD is a flexible classification system because a prediction of a category can be made on any subset of the features describing an observation. The system can also classify new observations that have missing data.

It is hypothesized that a PD classifier can meet most of the requirements discussed in Section 2.3 because it has the following advantages over existing classification methods: a) able to discover multivariate rules comprised of mixed mode features, b) transparent because it can reveal the significant patterns, discovered as a set of feature values, which contributed most significantly to the

classification, c) the ability to handle any underlying probability density and d) the ability to classify new observations that have missing data. It can easily handle discrete features. For instance observations made by a physician about a patient's symptoms are often discrete, unordered (nominal) variables that can be defined by a finite alphabet.

Chapter 3
MUP Features, MUP Data and Performance Measures

This chapter defines the features measured from MUP templates and describes the different MUP data sets used in this work. The MUP data sets are organized by the type of decomposition system used to establish MUP templates either DQEMG or Multi-MUP. The chapter ends by defining the type of measures used to evaluate the performance of different characterization systems.

3.1 Definition of MUP Features

The following definitions are from [34] unless otherwise noted. Refer to Figure 3.1 for a graphical depiction of many of the features defined below.

Duration (ms) is defined as the time between starting point (onset) and end-point of a MUP. These points are often determined by deviation from the baseline and exceeding a minimum slope criterion. The onset and end markers defined by automatic algorithms need to be inspected manually by a human operator and sometimes may need to be reset.

Spike duration (ms) is defined as the time between the first and last positive peaks of the MUP.

Amplitude (μV) of a MUP is the difference in voltage from the maximal negative to the maximal positive peak within the duration of a MUP.

Area (μV·ms) is calculated from the rectified MUP signal within the duration.

Thickness (ms) is defined by area-to-amplitude ratio (AAR) [35].

Size Index is a logarithmic function of thickness and amplitude that is related to the size and shape of a MUP [36, 37].

A *phase* is the part of a MUP that falls between baseline crossings and exceeds a minimal amplitude threshold. The number of phases is counted within the duration.

A *turn* is a local peak, either negative or positive in the MUP waveform. Peaks generated by noise are excluded by defining a turn as a peak that exceeds a minimum voltage change between successive peaks.

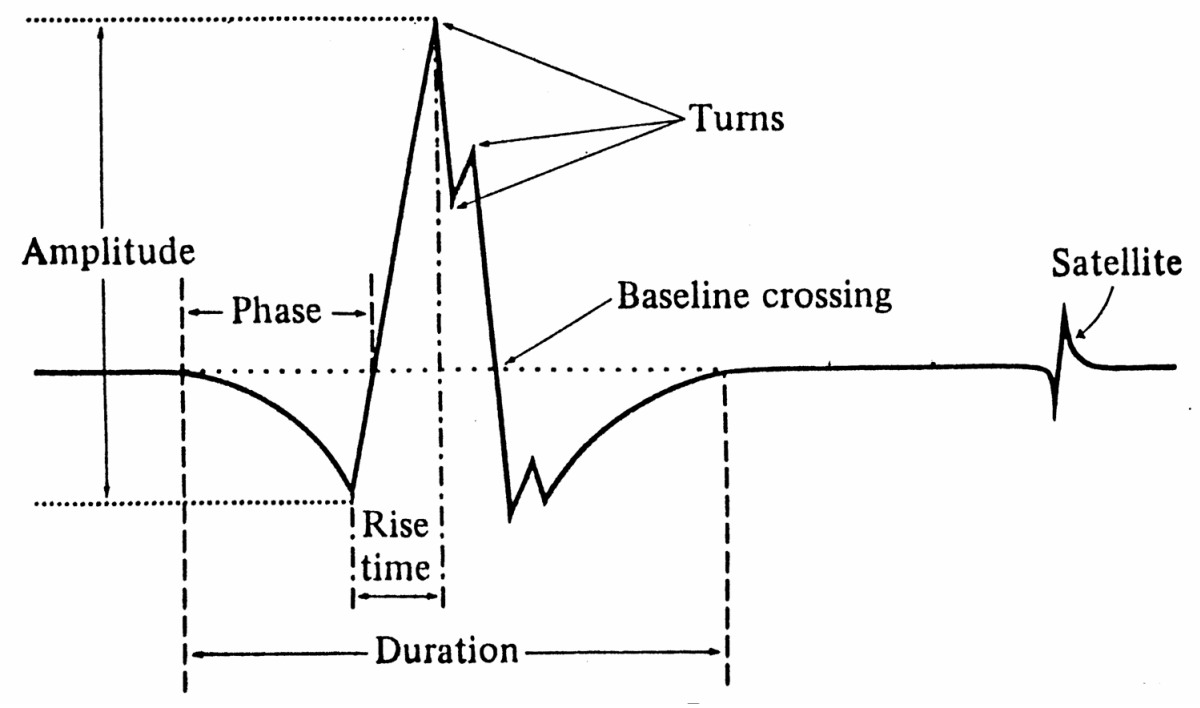


Figure 3.1 MUP Feature Depictions (From D.W. Stashuk, T.J. Doherty: "Normal Motor Unit Action Potential" in Neuromuscular Function and Disease , vol. 1, W. F. Brown, C. F. Bolton and M. J. Aminoff, Eds. Philadelphia, PA: Elsevier Science, 2002, pp. 291-310.)

3.2 DQEMG Decomposed MUP Data

MUP data described in this Section were estimated by decomposing EMG signals using DQEMG [7]. DQEMG typically finds 51 isolated MUPs produced by a single motor unit, aligns them using maximum slope, and uses a median trimmed average to form the MUP template. DQEMG uses a 25 μV threshold to define a turn and a phase has to have amplitude of at least 20 μV and duration of at least 240 μs. A revised Size Index (rSI) was used for MUP data decomposed by DQEMG and was calculated by $rSI = \log_{10}\left(amp_{(mV)}\right) + 1.8 \cdot \log_{10}\left(thickness_{(ms)}\right)$ [37].

3.2.1 Simulated Data

EMG signals were simulated using a physiologically-based model [38] to help examine the relationship between level of involvement and MUP characterization performance. The simulator was extended to also allow simulation of the affects of neuropathic and myopathic disorders [39]. To simulate a neuropathy, motor units are reorganized progressing from random motor neuron death to random re-innervation of orphaned fibers by nearby surviving motor neurons. To simulate inflammatory myopathy, a small percentage of randomly selected healthy fibers are "infected" and atrophied by a small fraction, and a smaller proportion are hypertrophied by a small fraction. This process is iterated by infecting additional fibers, and atrophying and hypertrophying the newly and previously infected fibers until the prescribed level of involvement is reached. A fiber is considered non-functioning (i.e. dead) when its diameter is below a critical threshold.

The simulator models the recruitment of motor units necessary to bring the level of force produced by a diseased muscle up to a prescribed percentage of the maximal voluntary contraction (MVC) force of a healthy muscle. Therefore, a simulated muscle that is normal, or with 25% or 75% motor unit loss at 10% MVC are each modeled as if producing close to the same force. EMG signals detected using a concentric needle at various intramuscular positions of several different muscles during approximately 7 to 10% MVC contractions were simulated and then decomposed into MUP templates. This method mimics the completion of several EMG studies across different individuals and includes levels of background MUP interference and noise typical of clinical studies. A comparison done by Hamilton-Wright and Stashuk [38] of the quantitative analysis of simulated versus real healthy decomposed MUPs shows good correspondence.

In total, 500 MUPs were extracted from simulated EMG signals of normal muscle, 500 from myopathic muscle and 500 from neuropathic muscle. The myopathic/neuropathic MUPs, in approximately equal proportion, were simulated to come from muscles with 25%, 50% and 75%

muscle-fiber/motor-unit loss. Each MUP was labeled as either normal, myopathic or neuropathic allowing comparison with the literature that use these labels [1, 10, 11] .

3.2.2 London Health Sciences (LHS) MUP Data

Control and neuropathic MUPs were sampled from the biceps-brachii and first dorsal interosseous (FDI) muscles. In total, 1649 MUPs from 54 muscles were sampled from 16 healthy control subjects (aged 27 ± 4 years) and 427 MUPs from 22 muscles were sampled from 14 patients, including 9 patients (aged 52 ± 12 years) with clinically probable or definite amyotrophic lateral sclerosis as defined by the revised El Escorial criteria [40] and 5 patients (aged 37 ± 11 years) with Charcot-Marie-Tooth Disease Type 1X confirmed via genetic testing.

A disposable concentric needle electrode (Model N53153; Teca Corp., Hawthorne, NY) was used to acquire EMG signals during 30 s voluntary isometric contractions performed at between 10% and 20% of each individual subjects' maximal voluntary contraction using DQEMG on a Neuroscan Comperio (Neuroscan Medical Systems, El Paso, Texas) with a bandpass of 10 Hz–10 kHz at a sampling rate of 31.2 kHz as previously described [41-43].

3.2.3 Rigshospitalet (Rigs) Data

A standard concentric needle electrode was used to acquire EMG signals for 11.2 seconds during voluntary contraction of various muscles and then amplified by a high impedance differential amplifier (DISA, 15C01) and analog band-pass filtered with high and low pass filters set at 2 Hz and 10 kHz. The EMG signals were sampled and digitized at a sampling frequency of 23437.5 Hz with 16 bit resolution (Motorola DSP56ADC16).

MUP data was acquired from biceps muscles to establish control data. MUP data was acquired from primarily biceps and vastus medialis muscles from patients with different types of myopathy (various forms of dystrophy, polymyositis and unknown myopathies) and ALS. MUP data was also

acquired from FDI, APB (Abductor pollicis brevis), and TA (Tibialis Anterior) muscles of patients. More information about this data is available in [44].

3.3 Multi-MUP Decomposed MUP Data

The data described in this section was obtained by decomposing EMG signals using the multi-MUP technique described previously [45, 46] where many different individual MUPs, detected at a single needle insertion point, that are deemed to belong to the same MUP train by a shape based classifier are averaged. The mean and SD of time points of the set of MUPs are calculated. The MUP template is calculated by excluding points that lie outside the ±1 SD in the time point average. The Multi-MUP system has many similarities with DQEMG. However, a drawback of the Multi-MUP system is its inability to capture MUPs with long risetimes according to Miki Nikolic who has had extensive lab experience with the system [44]. Another difference between Multi-MUP and DQEMG is the 50 μV [45] threshold for defining a turn compared to the 25 μV threshold used by DQEMG. Multi-MUP also defines size index (SI) as $SI = 2 \cdot \log_{10}\left(amp_{(mV)}\right) + thickness_{(ms)}$ [36].

3.3.1 University of Ljubljana External Anal Sphincter (EAS) MUP Data

Quantitative MUP data from a group of 86 (58 men) patients [46] (Podnar, et al. 2002) (called patient-sensitivity) was used to study sensitivity. Data from a group of 77 (49 female) control subjects [47] (Podnar. 2004) (called control-specificity) was used to study specificity, and data from a separate group of control subjects sampled from 64 control subjects [46] (Podnar, et al. 2002) (called control-reference) was used to establish normative thresholds. The patients had cauda equina or conus medullaris lesion. The control-specificity group consisted of subjects referred for minor pelvic floor dysfunction but whose examination showed no neuromuscular or other related disorders. The normative thresholds of the control-reference data were set at +/- 2 SD for the means and at the 5^{th}-

95th percentiles for outliers and were previously published [46] (Podnar, et al. 2002). The diagnosis of the subjects and patients was made as per standard clinical practice reported previously [45, 47-49].

Intramuscular EMG signals were detected using a concentric needle electrode and a commercial EMG system (Keypoint; Alpine Biomed Neurodiagnostics, Skovlunde, Denmark) with a bandpass of 5 Hz–10 kHz as previously described [50] (Podnar and Vodusek. 1999). Using the multi-MUP technique described previously (Podnar, et al. 2002, Stålberg, et al. 1995) individual MUP waveforms were estimated and their feature values calculated [46].

3.3.2 University of Ljubljana Biceps MUP Data

EMG signals were acquired from 33 biceps muscles of healthy subjects and 30 biceps muscles from patients with a genetic diagnosis of facioscapulohumeral muscular dystrophy (FSHD) using a standard concentric EMG needle, and a commercially available EMG system (Keypoint; Medtronic Functional Diagnostics, Skovlunde, Denmark) with standard settings (filters: 5 Hz to 10 kHz) [51]. Sampling of a muscle continued until at least 20 different MUPs per muscle were decomposed using multi-MUP analysis.

3.4 Distribution of MUP Thickness Feature among Categories

The distribution of MUP thickness values for each category for the data used in this thesis was estimated using a Parzen window technique using a Gaussian kernel. The distributions were used to represent the separation of the data across the different categories to gauge the difficulty in accurately characterizing the data. Categories that are more separated should be easier to characterize accurately than categories that are less separated. Thickness was chosen as the feature to represent separation because it is thought to be highly discriminative [35, 52]. Other features showed similar distributions per category and data set.

Figures 3.2 and 3.3 show the estimated distributions of thickness for the simulated and Rigs data respectively. They show that myopathic data has the lowest mean thickness followed by normal and then neuropathic data. This is expected because myopathic data tends to have the thinnest MUPs followed by thicker normal and then neuropathic MUPs. Figure 3.2 shows that the variance of myopathic and neuropathic data is larger than that of normal which is expected. Figure 3.4 and 3.5 show the estimated distributions of thickness for the LHS and EAS data respectively where the normal category has a lower mean and variance than the neuropathic category.

The mean and standard deviation of thickness values for the Rigs, LHS, and the Ljubljana biceps data (LJUB-Biceps) shown in Table 3.1 are similar for respective categories. The mean and standard deviation of thickness values for the Rigs data compared to the LHS data shown in Table 3.1 are slightly different.

All of the data sets have a great deal of overlap in the distributions of thickness values among the different categories as expected according to [18] but there is some separation in the data among the categories. However, the EAS data does not have a great deal of separation between its categories as shown by how closely the distributions align in Figure 3.5 and the small distance value of the EAS data compared to the other data sets as shown in Table 3.1. All of the other data sets have more separation between categories than compared to the EAS data. The LHS and Ljubljana-biceps data had very similar distances while the Rigs data had the greatest separation of all the MUP data sets as shown in Table 3.1. It is expected that the EAS data will have lower characterization accuracy and the Rigs data set will have higher characterization accuracy than the other MUP data sets based on distance between categories of their thickness distributions

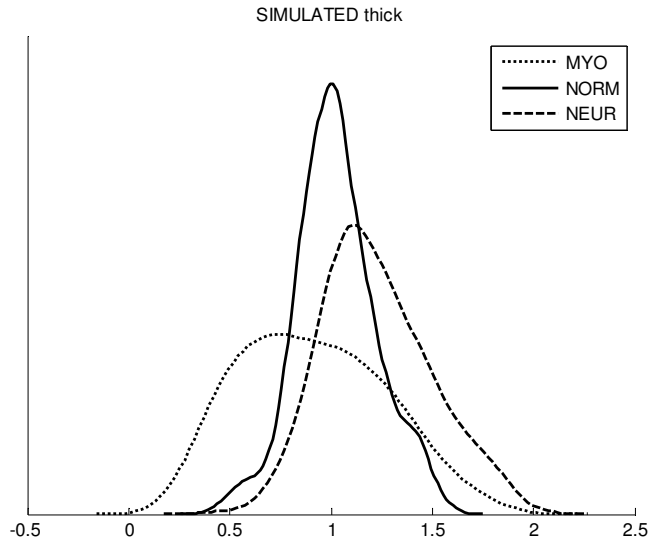


Figure 3.2 Estimated Distribution of Thickness - Simulated Data

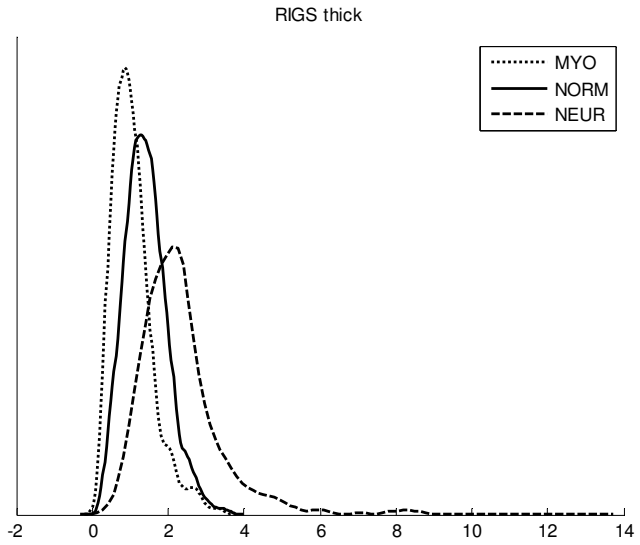


Figure 3.3 Estimated Distribution of Thickness - Rigs Data

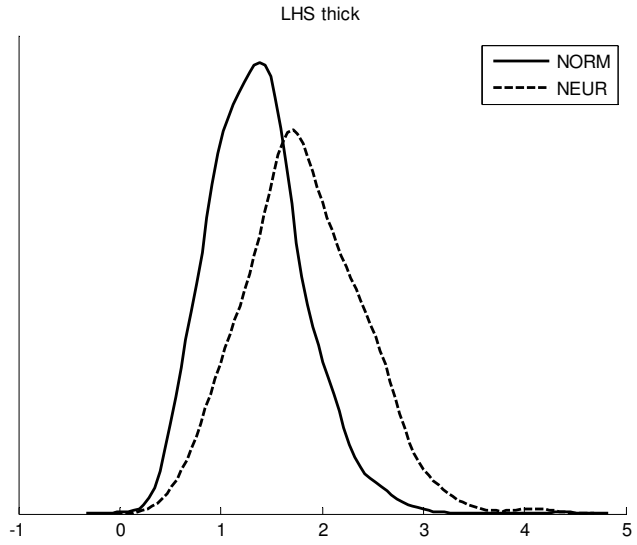


Figure 3.4 Estimated Distribution of Thickness - LHS Data

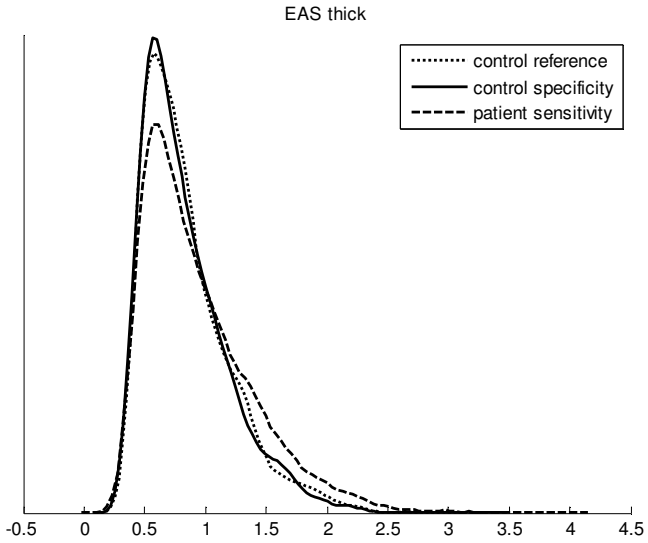


Figure 3.5 Estimated Distribution of Thickness - EAS Data

Table 3.1 Mean, Standard Deviation and Distance of Thickness Distributions

	MYO		NORM		NEUR		
DATA	mean	SD	mean	SD	mean	SD	Distance
SIM	0.91	0.35	1.03	0.20	1.23	0.27	0.38
RIGS	1.06	0.54	1.41	0.55	2.36	1.25	0.52
LHS			1.36	0.48	1.82	0.58	0.43
EAS			0.85	0.37	0.94	0.46	0.11
LJUB-Bicep	1.03	0.45	1.40	0.45			0.42

The distance shown in the last column in Table 3.1 for the two-category distributions was calculated by $D = (m_1 - m_2)/(\sigma_1 + \sigma_2)$ where m_i is the mean for category *i* and σ_i was the standard deviation of category *i.* Distance for the three-category distributions was calculated by taking the average across all combinations of two-category distances.

3.5 Performance Measures

When analyzing the EAS, LHS and Ljubljana biceps data that had only two categories, sensitivity and specificity were the terms used to describe per category accuracy. Sensitivity for the EAS and LHS data was defined as the total number of muscles characterized as neuropathic divided by the total number of 'true' neuropathic, i.e. patient muscles. Sensitivity for the Ljubljana biceps data was defined as the total number of muscles characterized as myopathic divided by the total number of 'true' myopathic, i.e. patient muscles. Specificity was defined as the total number of muscles characterized as normal divided by the total number of 'true' normal, control muscles. Accuracy was defined as the average of sensitivity and specificity. The traditional definition of accuracy *(true negatives + true positives)/(all muscles tested)* was not used because it is biased towards the category that has the largest number of test muscles to be characterized. The traditional accuracy measure would be skewed by an unequal proportion of controls to patients. For instance, the traditional accuracy measure would under weigh any results for the patients if there were fewer patients than controls.

The term sensitivity-specificity deviation (SSD) was defined as:

$$SSD = \sqrt{\left((A - Sens)^2 + (A - Spec)^2\right)/2} \qquad (3.1)$$

where A is accuracy - the mean of specificity and sensitivity.

$Sens$ is sensitivity, and

$Spec$ is specificity,

For data sets with three categories, i.e. Rigs and simulated data, overall accuracy A was the mean across per-category accuracies, i.e. A_k is the accuracy for category y_k – the number of observations of category y_k that were classified as y_k divided by the total number of observations with 'true' category y_k. The SSD for three categories is then:

$$SSD = \sqrt{\left((A - A_{myo})^2 + (A - A_{norm})^2 + (A - A_{neur})^2\right)/3} \qquad (3.2)$$

where A_{myo}, A_{norm} and A_{neur} are the accuracies for the myopathic, normal and neuropathic categories respectively.

SSD was used to determine how well a characterization method maximized/balanced both specificity and sensitivity.

Chapter 4
MUP Characterization

This chapter describes the Pattern Discovery method in detail and then applies it for the characterization of MUPs. The PD method was evaluated and compared to other pattern recognition techniques using the criteria for a CDSS described in Section 2.3.

4.1 Pattern Discovery Based Classification.

A MUP can be characterized using Pattern Discovery (PD) by discovering rules in a set of training data. Further information about this PD method can be found in [30-32, 53]. Given a training data set with *N* MUP samples, a MUP $\mathbf{X}$ is described by *M* features $\mathbf{X} = \{x_1, \cdots, x_M\}$ where each feature $x_i \quad 1 \le i \le M$ is a random variable. Each feature can take on a value from its discrete alphabet $\alpha_i = \{\alpha_i^1, \ldots, \alpha_i^{m_i}\}$ where m_i is the number of characters in the alphabet of the ith feature. A primary event occurs when a single feature x_i takes on a value from α_i. The p^{th} primary event is denoted by $[x_i = \alpha_i^p]$ or simplified to x_{ip}. X^r is an r^{th} order event that contains r primary events. A set of category labels $\{y_{1,} \cdots, y_k, \cdots, y_K\}$ contains K labels and $1 \le k \le K$. The number of features, including a category label, determines the order of an event. An example of a 4th order event would be a MUP from a neuropathic muscle, i.e. labeled neuropathic, with high amplitude, high duration and a high number of turns. For continuous data, the range of feature values can be divided into pre-defined intervals, i.e. quantized, and each interval labeled in a natural language as is often used in the clinic: e.g. small, medium, or large. If the number of observations (denoted by o_{X^s}) of an event is significantly higher or lower than the number of expected observations (denoted by e_{X^s}) assuming randomly distributed data then an event is said to be a pattern when it passes a test of statistical

significance using the adjusted residual d_{X^s} [30] and the number of expected observations exceeds a threshold (usually $e_{X^s} = 5$) to ensure statistical robustness. The expected number of observations is given by:

$$e_{x^s} = N \cdot \prod_{x_{ip} \in x^s} \Pr(X_i = \alpha_i^p), \qquad (4.1)$$

where *N* is the number of MUPs and $\Pr(X_i = \alpha_i^p) = o_{x_{ip}} / N$

Equation 4.1 calculates the expected number of observations based on the NULL hypothesis that the feature values comprising an event are independent. An event is called a pattern when the NULL hypothesis is rejected by testing using adjusted residual. The adjusted residual is calculated by first finding the standardized residual.

$$z_{X^s} = \frac{o_{X^s} - e_{X^s}}{\sqrt{e_{X^s}}} \qquad (4.2)$$

The standardized residual is considered to have a Gaussian distribution when the asymptotic variance of z_{X^s} is close to one, otherwise the standardized residual has to be adjusted to a Gaussian distribution with a variance of one. This adjusted residual is expressed by:

$$d_{X^s} = \frac{z_{X^s}}{\sqrt{v_{X^s}}}, \qquad (4.3)$$

where v_{X^s} is the maximum likelihood estimate of the variance of z_{X^s}

A positive pattern occurs if $d_{X^s} > 1.96$ at 95% confidence level.

A negative pattern occurs if $d_{X^s} < -1.96$ at 95% confidence level.

If the pattern contains a category label then it is a rule. Because they contain category labels and are significant based on a statistical test, rules can be used for characterization. The discriminatory power

of the l^{th} r^{th} order rule, X_l^r, can be determined by its weight of evidence (WOE), which is the odds of a sample MUP, which has a subset of feature values that match the feature values of X_l^r, belonging to category y_k versus not belonging to y_k.

$$WOE = \log_2 \frac{P(X_l^r \mid MUP = y_k)}{P(X_l^r \mid MUP \neq y_k)} \qquad (4.4)$$

where:

MUP = y_k, MUP is detected in a muscle of category y_k.

MUP ≠ y_k, MUP is detected in a muscle not of category y_k.

P(X | Y) the probability of X occurring given that Y has occurred – a conditional probability.

X_l^r an r^{th} order pattern, where l is an index indicating the l^{th}, r^{th} order pattern.

Note that the range of (4.4) is $-\infty < WOE < +\infty$. A rule labeled y_k with positive WOE provides support for the category y_k. A rule supporting a label other than y_k will have a negative WOE and provides refutation for the category y_k. The strength of the support (or refutation) is proportional to the absolute value of the WOE. An r^{th} order rule where $r < M$ is called a component rule. The union of disjoint component rules of the same category, i.e. the union of the set of component rules with no overlapping features, discovered in a MUP is called a compound rule and is denoted by x_k^*. There is one compound rule per category per MUP and the category is denoted by the subscript k in x_k^*.

The weight of evidence for a category for a MUP under test is the sum of the WOEs of each component rule that belongs to the compound rule associated with the MUP under test. The total

WOE provides either support or refutation of category y_k. The PD classifier calculates K weights of evidence (one for each category) for each MUP under test.

4.2 Evaluation of MUP Characterization Methods

The performance of several classifiers used to characterize the simulated MUP data (Section 3.2.1) and LHS MUP data (Section 3.2.2) were evaluated.

4.2.1 Other Classifiers Considered for MUP Characterization

The classification performance of three other classifiers was compared with the PD classifier: two discrete classifiers – Weka J48 DT and Weka NB; and one continuous – LDA. Although not as transparent as the other classifiers, LDA was chosen for comparison because it was previously used for MUP classification [1, 11]. DT and NB classifiers were chosen because they are commonly used in medical decision support, are regarded as being transparent [14, 26], and can handle nominal data. The DT and NB classification error rates were determined using the algorithms of the Weka explorer system [24]. A Matlab function, RAFisher2CDA [54], was modified to calculate the Fisher discriminants used for LDA classification.

4.2.2 Training and Testing

Thirty trials were conducted using the PD, LDA, DT, and NB classifiers with the simulated MUP data partitioned into different sets of training and test data. There was a complete separation of training and test data for each trial. Three hundred samples from each category were randomly chosen for training per trial for a total of 900 training samples. Each trial used the same training and test data for the different classification methods. After the training data was chosen, the remaining data was used to test the classifiers to establish classification performance. A mean classification error rate and standard deviation was calculated to determine classifier performance. This process was repeated for

the LHS MUP data except the LHS data consists of two categories and so 600 training samples per trial were used.

The classifiers used the following MUP features: amplitude, duration, phases, turns and thickness otherwise known as area-to-amplitude ratio (AAR). Except for LDA, which requires continuous data, the MUP data was quantized into intervals before being used for classification. The discrete classifiers (PD, DT and NB) were all presented with the same quantized training and testing data sets. This prevented NB and DT from using their own quantization methods so that the effects of quantization on classification accuracy were not a factor in the comparison. Using only three intervals keeps the number of rules and the size of the decision tree to a reasonable size; and features can be easily translated into linguistic labels such as low, medium, and high which is conceptually consistent with clinical practice. As well, three intervals help to simplify the visual patterns that explain the results leading to diagrams that are easily recognized and understood by clinicians. The natural logarithm of amplitude was taken before determining the linear discriminant functions to minimize its skewness. Prior to quantization none of the features were transformed for PD, DT, and NB classification.

The classification error rate of using a single MUP to characterize a muscle was examined. The per category error rate was defined as the number of MUPs that incorrectly predicted the muscle category divided by the total number of MUPs detected from muscles of that category.

4.3 Results LHS MUP Data

PD had an average error rate (and standard deviation) of 30.3% (1.8%), LDA 29.0% (2.1%), DT 30.1% (2.1%) and NB 29.8% (1.8%) across thirty trials. All four characterization methods had similar error rates with no statistically significant differences between each other according to the Tukey post hoc test at a significance level of 0.05. Table 4.1 shows NB had the highest sensitivity and LDA had

the highest specificity and accuracy. However, PD had a significantly lower SSD compared to the other methods. J48 had the next best SSD, which was about 2 times the SSD of the PD method.

4.4 Results Simulated MUP Data

PD had an average error rate (and standard deviation) of 24.3% (1.2%), LDA 23.1% (1.4%), DT 24.6% (1.2%) and NB 27.0% (1.4%) across thirty trials. A Tukey post hoc test showed that LDA had a significant improvement at the 0.05 level compared to the other classifiers. PD and DT did not have statistically significant differences between each other according to the Tukey post hoc test at a significance level of 0.05. The Tukey post hoc test between the other classifiers and NB showed that it had a significant decrease in performance at the 0.05 level.

Tables 4.2 and 4.3 show the confusion matrix for PD and LDA simulated MUP data classifications respectively. The confusion matrices shown were determined as an average of the confusion matrices across the thirty trials. The Tables show that the per category error rates for simulated normal and neuropathic are similar with a dramatic improvement for simulated myopathic data.

If only normal and neuropathic simulated MUP data error rates are compared, Table 4.2 shows an average error rate across two categories of 29% for the PD classifier and Table 3 shows a two category error rate of 28.8% for the LDA classifier, which are very similar to the error rates of the LHS data of 30.3% and 29.0% respectively. Note that the LHS data did not include any myopathic data.

Figure 4.1 shows that as the level of simulated involvement increased, the classification error rate decreased for MUPs detected from simulated myopathic and neuropathic muscles. Neuropathic MUPs had error rates of 48%, 30%, and 21% for levels of involvement of 25%, 50%, and 75% respectively. Myopathic MUPs had error rates of 22%, 13%, and 7% for levels of involvement of 25%, 50% and 75% respectively.

Table 4.1 Sensitivity/Specificity and Accuracy of Characterization of LHS MUP Data

	Sensitivity	Specificity	Accuracy	SSD
PD	71.7%	67.8%	69.7%	1.9%
LDA	65.8%	76.2%	71.0%	5.2%
J48	66.3%	73.5%	69.9%	3.6%
NB	75.7%	64.8%	70.2%	5.4%

The results in Table 4.1 are the mean sensitivity, specificity, and accuracy across thirty trials for the LHS MUP data. The SSD is based on the means shown in Table 4.1. The training data for each trial consisted of three hundred randomly chosen samples from each category for a total of 600 samples. After the training data was chosen, the remaining data was used to test the classifiers to establish classification performance. Each trial used the same training and test data across the different classification methods.

Table 4.2 Pattern Discovery Confusion Matrix (simulated MUP data)

	Classified as			
True Class	Myopathic	Normal	Neuropathic	per class error rate
Myopathic	170	20	10	15%
Normal	20	144	36	28%
Neuropathic	21	39	140	30%

The PD confusion matrix shows that myopathic simulated MUPs had the lowest per-category classification error rate of 15%. The neuropathic MUPs had the highest per-category error rate at 30%. This confusion matrix shows the expected per-category distribution of errors. The probability of error that a MUP detected from a myopathic muscle is misclassified as normal is 2 times greater than being misclassified as neuropathic (20/200 versus 10/200). A similar trend appears for neuropathic data where the probability of error of a MUP being misclassified as normal is about 2 times greater than being misclassified as myopathic (39/200 versus 21/200).

Table 4.3 LDA Confusion Matrix (simulated MUP data)

	Classified as			
True Class	Myopathic	Normal	Neuropathic	per class error rate
Myopathic	176	20	4	12%
Normal	17	138	45	31%
Neuropathic	15	38	147	27%

The LDA confusion matrix shows that there was an increase in the error rate for normal classifications from 28% to 31% and a decrease in error rate from 30% to 27% for neuropathic classifications compared to the PD classifier. The confusion matrix for LDA shows the expected distribution for incorrectly classified MUPs detected from muscles with disorders is 5 times greater for normal than myopathic and about 2.5 times greater for normal than neuropathic.

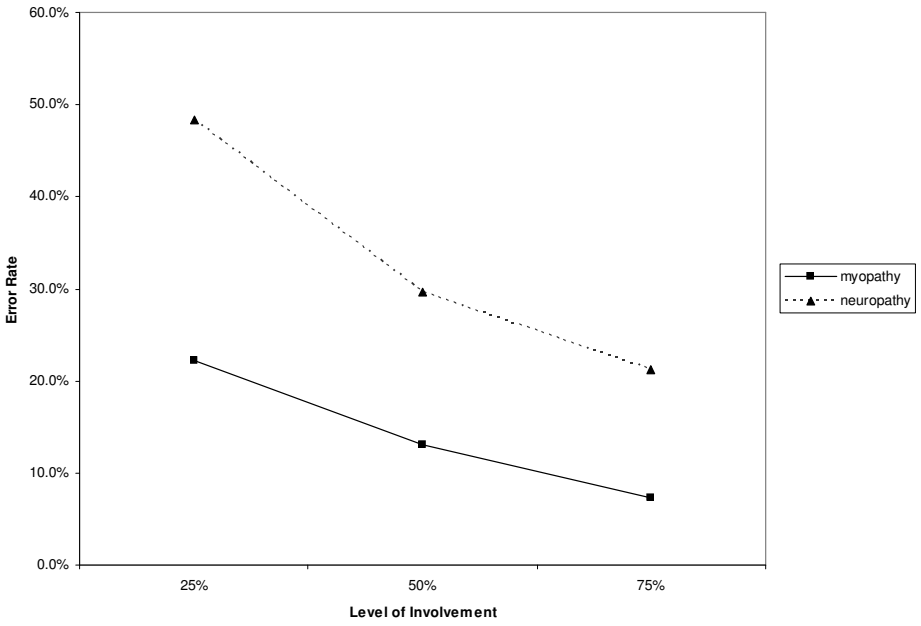


Figure 4.1 Error Rate vs. Level of Involvement for Simulated MUPs

4.5 Discussion MUP Characterization

Table 4.4 summarizes a comparison of the studied classifiers for four of the identified requirements based on the study results and consideration of the known properties of each classifier. PD has an advantage with respect to transparency. All four methods have similar accuracy. PD, DT and NB can handle mixed mode data types while LDA cannot because of its inability to handle nominal data. All four classifiers can produce a numeric value for characterization. In addition, while the computational effort required for training for each method can be significant, it is done offline prior to clinical use and only needs to be done once. The computational effort for characterization expended by each method is similar and not consequentially different from current QEMG examinations.

Figures 4.2 to 4.4 are example PD characterizations of three different LHS MUPs and Figure 4.5 is an example characterization of a simulated MUP. The reporting of the component rules that support each characterization demonstrates the transparency of PD characterization. Furthermore, the rules discovered by PD are consistent with currently used diagnostic criteria [34, 35, 55, 56] (see Table 4.5). These MUP characterizations demonstrate that the PD classifier is capable of capturing knowledge consistent with current practice and expressing it in transparent, understandable terms. The knowledge is captured while simultaneously considering multiple feature values and expressed using common clinical terminology.

PD characterization provides a transparent explanation of the set of features weighted by their contribution, using both negative and positive rules that refute or support characterization towards the various categories. The system is able to objectively calculate the WOE that a MUP was detected from a normal or neuropathic muscle. This reduces the mental workload of the clinician, which should reduce the number of errors in both individual MUP characterizations as well as overall muscle characterizations. This leads to the expectation that examining clinicians will make fewer errors.

The error rate of the LDA and PD method based on normal and neuropathic LHS MUP data is about 25% lower than the error rate achieved in Pfeiffer's work [1]. Pfeiffer's data [1], collected two years apart, was not done consistently. It is not expected that PD will achieve better accuracy than LDA. If PD were tested using the data set in [1] performance similar to LDA would be expected. PD's advantage however is transparency.

DT, NB and PD based classifiers are suitable for use in MUP characterization; however, PD has an advantage over DT and NB based on its transparency. The low error rate of the LDA method helps to confirm Pfeiffer's conclusion that LDA classification is an accurate method [1]. However, LDA is not very transparent and cannot easily handle nominal data types – an important requirement in extending a clinical decision support system to handle other clinical observations. DT classifiers use an error reduction based training algorithm and therefore provide rules that are used to minimize classification error. This can lead to over-fitting problems. NB and DT methods, as mentioned previously, cannot find relationships among multiple features. PD rules are determined through observation of statistically significant relationships in the training data without considering classification accuracy and capture the dependencies among features using hypothesis testing. Although accuracy across the four methods did not differ significantly for the LHS MUP data, PD had a significant advantage in its ability to maximize both sensitivity and specificity. A clinic with unknown "costs" for false negative and false positive characterizations would favour a characterization method that maximized both sensitivity and specificity.

The characterization methods had similar error rates for LHS MUP data, and simulated normal and neuropathic MUPs. This suggests that the simulator can be useful for studying the relationships between level of involvement and characterization performance for normal and neuropathic MUPs. However, the lower error rates achieved by the LDA and PD methods for simulated myopathic versus neuropathic data may be an artifact of the simulator. The simulator does not deal with changing

muscle density or structure. A separate study using clinical data from both inflammatory and non-inflammatory myopathic disorders is needed. Overall the error rate for the simulated MUP data was lower than the LHS MUP data because of the excellent performance of the simulated myopathic data. The similar performance of the methods for characterizing normal and neuropathic categories is consistent with the similarity in distance between the distribution of thickness values for the normal and neuropathic categories of the LHS and simulated MUP data as discussed in Section 3.4. If only the normal and neuropathic categories of the simulated MUP data were compared to the LHS data then the distance between these categories was 0.43 for both the simulated and LHS MUP data calculated using data from Table 3.1.

Although the results shown by Figure 4.1 are for simulated MUP data, they are consistent with the expectation that as a disease affects a greater portion of a muscle the probability of detecting MUPs that reflect the effects of the disorder increases thus reducing the number of errors made when categorizing MUPs detected from muscles with higher levels of involvement of a disorder. On the other hand, the reduced probability of detecting a MUP that reflects the affects of a disorder on a muscle during the early stages of involvement suggest that further development of QEMG methods, to combine MUP characterizations and/or use other QEMG based features, is needed to improve both the specificity and sensitivity with which a muscle can be characterized. Chapter 5 studies two methods for combining MUP characterizations using the LHS and other MUP data sets described in Chapter 3.

Table 4.4 Summary of Classifiers' Fit to Requirements

Requirement	PD	LDA	DT	NB
Transparency	∗∗∗∗	∗∗	∗∗∗	∗∗∗
Accuracy	∗∗∗	∗∗∗	∗∗∗	∗∗∗
Mixed –Mode Data	∗∗∗∗	∗∗	∗∗∗∗	∗∗∗∗
Numeric Characterization	∗∗∗∗	∗∗∗∗	∗∗∗∗	∗∗∗∗

Number of stars: Four – excellent, three – good, two – fair, one – poor. Accuracy ratings are based on the error rate discussed in the results. Transparency is based on the classifier's ability to report strength of support/refutation of a subset of features. Ability to handle mixed mode data and produce a numeric characterization scores are based on examining the methodology of the classifiers.

4.5.1 Conclusions

Unlike the other classifiers investigated, the PD classifier is able to explain itself by reporting the sets of feature values, along with a strength-of-evidence measure, supporting or refuting its characterizations. This work has demonstrated that the PD classifier meets the requirements for normal and neuropathic MUP characterization through its abilities to report its characterizations in a transparent manner, handle mixed mode data, discover dependencies among features, provide numerical characterization values, and achieve a similar level of characterization accuracy as state of the art classification methods. The PD classifier shows promise as a clinically useful method of providing numerical inputs to the next step of the interpretation phase of a QEMG examination (muscle characterization based on more than one MUP).

The PD classifier succeeds in interpreting the information extracted from quantitative MUP analysis and transforming it into knowledge that is consistent with current clinical practice. This demonstrates that the transparency provided by the PD classifier can be valuable for the capture and expression of knowledge useful to a clinician.

Characterization of the LHS and simulated, normal and neuropathic, MUP data had very similar error rates. These results provide encouragement to develop and evaluate a PD method for quantifying the level of involvement of a neuromuscular disorder – ultimately fulfilling one of the roles for future QEMG examinations envisioned by Swash [57].

Chapter 5 will focus on evaluating methods that combine MUP characterizations obtained using PD into muscle characterizations.

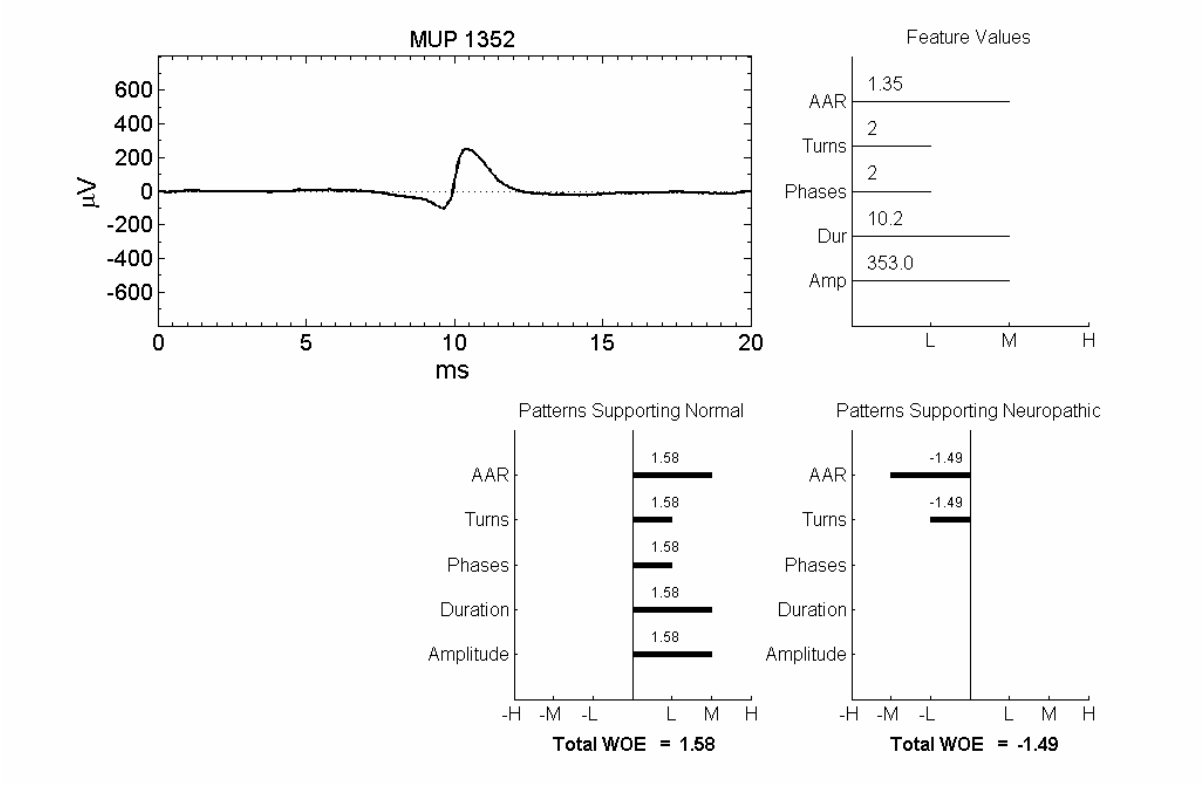


Figure 4.2 Evidence Supporting Characterization of LHS MUP 1352 as normal.

The upper lefts of Figures 4.2-4.5 show MUP templates each of which is the median-trimmed average of 51 isolated MUP firings deemed by DQEMG to belong to the same MUP train. Each of the MUP templates is displayed using the same scale. The y scale is in microvolts and the x scale is milliseconds. The upper right box of each Figure shows the result of quantization of the five MUP features into a low, medium or high quantization interval along with their continuous values prior to quantization. For instance in Figure 4.2 MUP 1352 is quantized as follows: thickness (AAR) of 1.35 is medium, 2 turns are low, 2 phases are low, duration of 10.2 ms is medium and amplitude of 353 μV is medium. The next two bar plots underneath show the component rules providing evidence for or against the MUP being detected from a normal or neuropathic muscle respectively. Features belonging to the same component rule are indicated by using the same line style (e.g. solid, dashed, dotted etc.). A component rule is positive if the bars go to the right from the vertical line and a component rule is negative if the bars are left of the vertical. The x-axis for a negative rule is also

marked with a negative sign for low, medium and high to emphasize that these are rules refuting the characterization.

The bar graph marked as 'Patterns Supporting Normal' shows that all of the features formed a single component rule since they are all solid leading to a WOE of 1.58 that MUP 1352 was detected from a normal muscle. The bar graph marked as 'Patterns Supporting Neuropathic' shows a single component rule for neuropathy. The component rule indicated by the solid line refuting neuropathic is composed of medium thickness (AAR), and low turns and has a WOE of -1.49. Since the WOE for normal has the highest value – this MUP was characterized as being detected from a normal muscle.

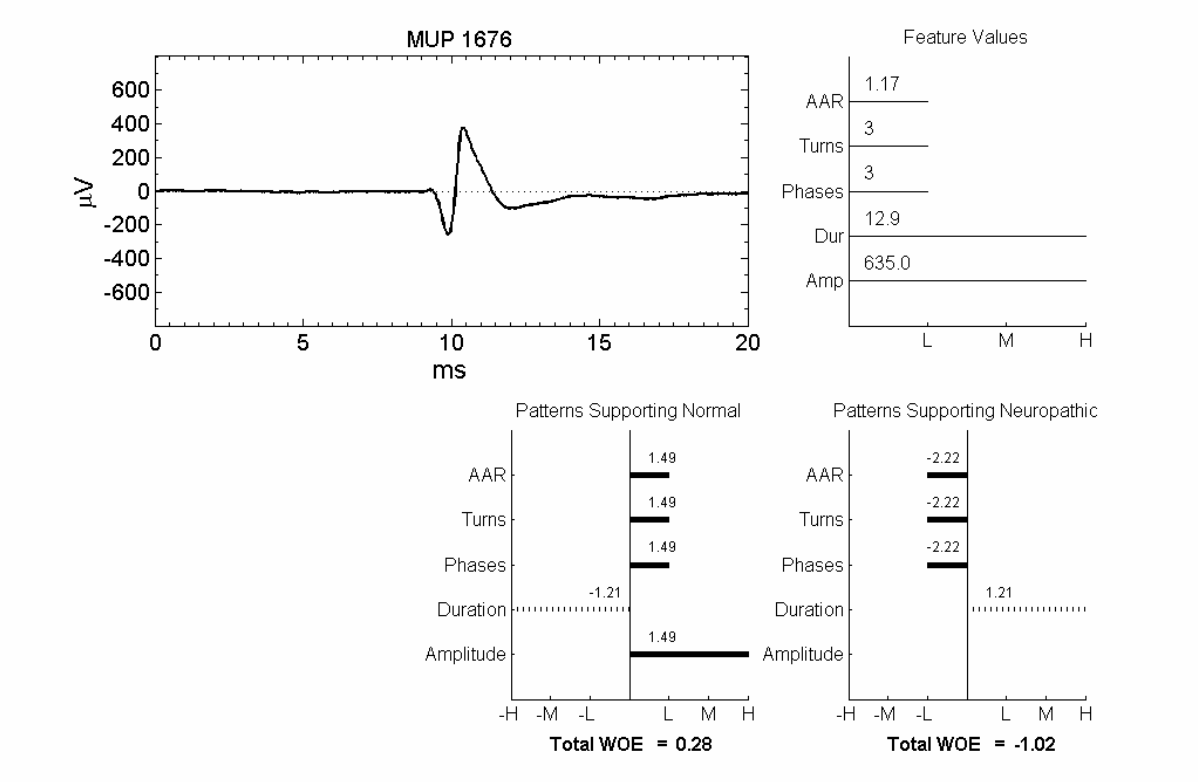


Figure 4.3 Evidence Supporting Characterization of LHS MUP 1676 as Normal.

The bar graph in Figure 4.3 titled as 'Patterns Supporting Normal' shows two component rules making up a compound rule for normal – one rule as a solid line, the other as a dashed line. The component rule indicated by the solid line supporting normal is composed of low thickness (AAR), low turns, low phases and high amplitude. The other component rule indicated by a dashed line is composed of high duration and refutes normal with a WOE of -1.21. In the 'Patterns Supporting Neuropathic' bar graph there are two component rules. The component rule indicated by the solid line with low thickness (AAR), low turns, and low phases has a negative WOE of -2.22. The other component rule indicated by a dashed line is composed of high duration and has a WOE of 1.21. In total, the WOE for being detected from a normal muscle is less than one at 0.28, and from a neuropathic muscle -1.02. The spread between the highest WOE and the second highest WOE for MUP 1676 is 1.3. This indicates a slightly lower level of confidence towards a classification for normal of MUP 1676 than MUP 1352 whose WOE spread was 3.07.

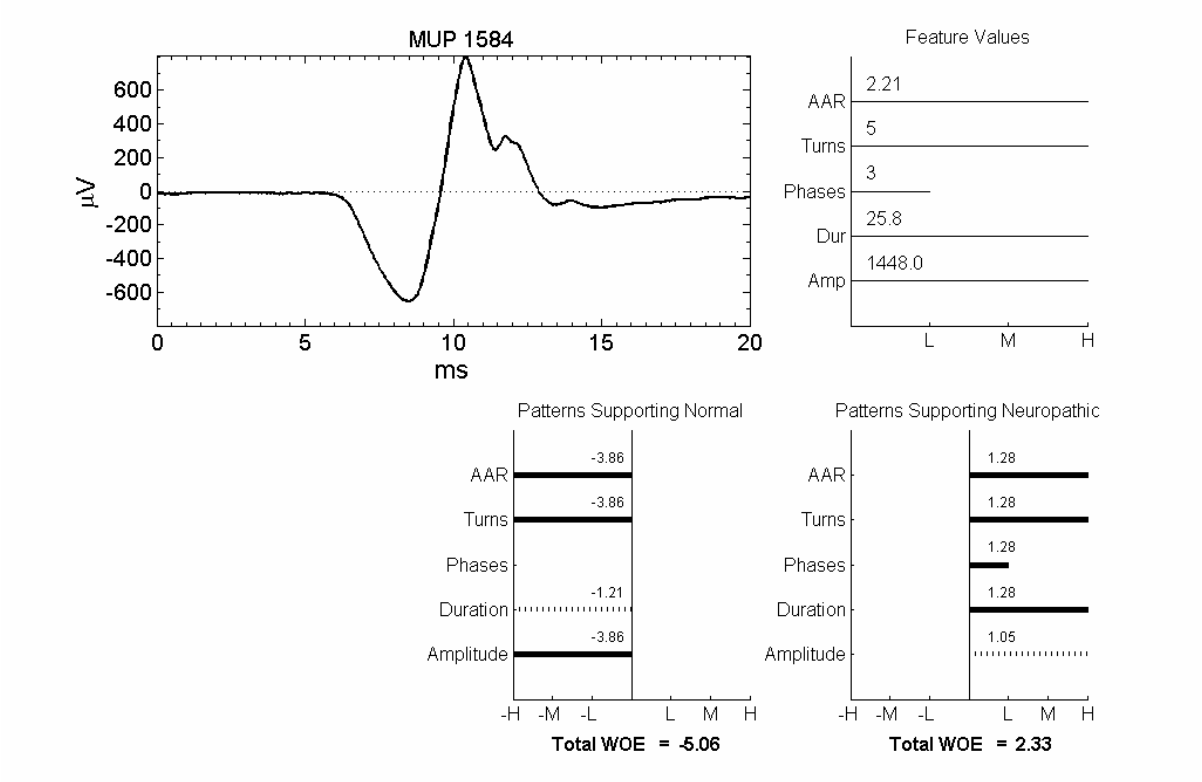


Figure 4.4 Evidence Supporting Characterization of LHS MUP 1584 as Neuropathic.

The bar graph in Figure 4.4 titled as 'Patterns Supporting Neuropathic' shows a compound rule comprised of two component rules that provide evidence that MUP 1584 was detected from a neuropathic muscle. A component rule (dashed line) of high amplitude provides a WOE of 1.05 and a component rule (solid line) of high thickness (AAR), high turns, low phases, and high duration provides a WOE of 1.28. In total, the two component rules provide WOE of 2.33 that this MUP was detected from a neuropathic muscle.

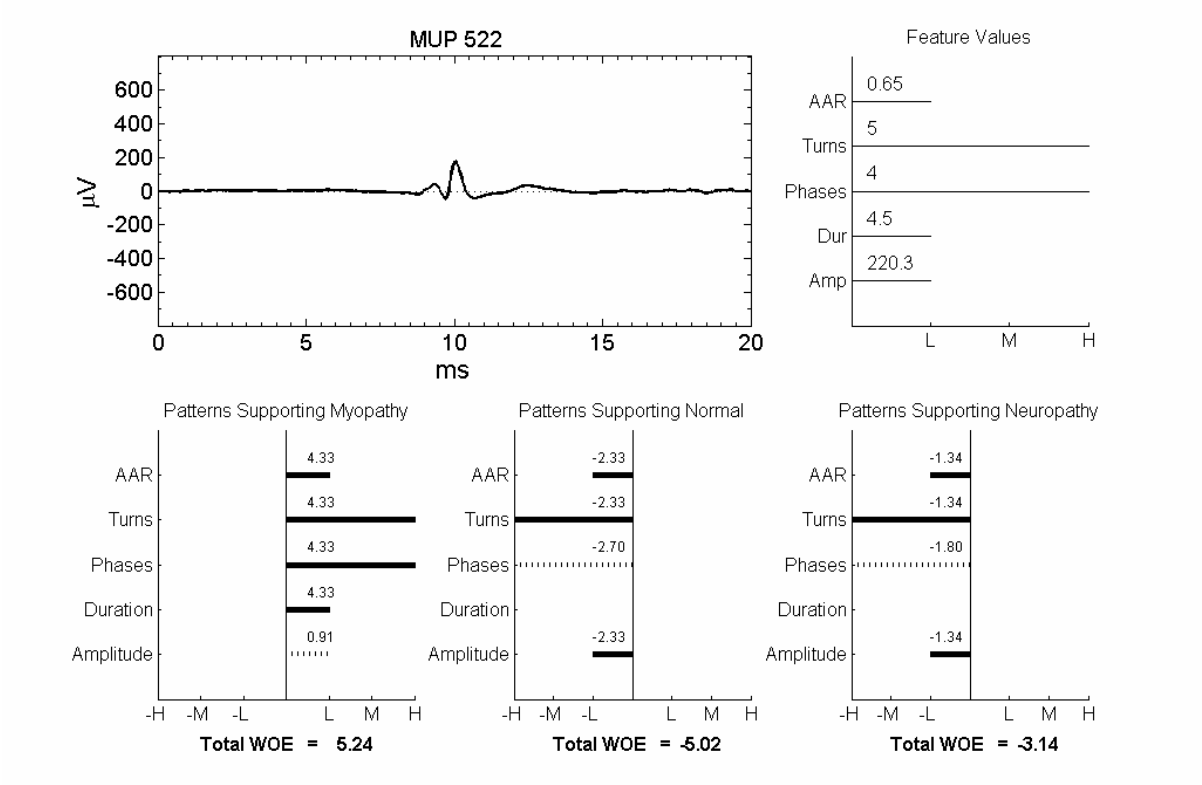


Figure 4.5 Evidence Supporting Characterization of Simulated MUP 1676 as Myopathic

The bar graph in Figure 4.5 titled as 'Patterns Supporting Myopathic' shows that the WOE of the first component rule indicated by a solid line of low thickness (AAR), high turns, high phases and low duration plus the WOE from the second component rule (dashed line) of low amplitude provides a total WOE of 5.24 that MUP 522 was detected from a myopathic muscle. The compound rules supporting normal and neuropathic both have a large spread (10.26 and 8.16 respectively) below the WOE value for myopathic; an indication of high confidence that this MUP was detected from a myopathic muscle.

Table 4.5 Current Clinical Criteria for MUP Characterization

#	Rule Description	Indication of	Source
1	decreased duration, decreased amplitude, polyphasic	myopathy	[56]
2	decreased AAR	myopathy	[35]
3	increased duration, increased amplitude, polyphasic	neuropathy	[56]
4	increased amplitude (and/or area) with normal or increased AAR	neuropathy	[35]
5	increased amplitude, high number of turns, may not be polyphasic	neuropathy	[34]

LHS MUP 1352 (Figure 4.2) with medium duration and amplitude, low number of turns, low phases and medium thickness (AAR) was characterized by the PD classifier as being normal. LHS MUP 1584 (Figure 4.4) with high duration and amplitude, low phases and high turns and high thickness (AAR) was characterized by the PD classifier as being neuropathic agreeing with rules 4 and 5. Simulated MUP 522 (Figure 4.5) with low duration and amplitude, high numbers of turns and phases and low thickness (AAR) was characterized by PD as being myopathic agreeing with rules 1 and 2.

Chapter 5
Muscle Characterization

Individual MUP characterizations lack sufficient information to accurately characterize a muscle; as such, multiple MUP characterizations need to be combined to provide an accurate muscle characterization. In this chapter we examine two methods for calculating muscle characterization measures: arithmetic mean and Bayes' rule. Given a set of MUPs detected from a muscle under examination, the information provided by each MUP can be combined into a Probabilistic muscle characterization. An example of a probabilistic muscle characterization is a 79% probability the muscle is neuropathic and a 21% probability the muscle is normal given the set of MUP characterizations based on the MUPs detected from that particular muscle.

5.1 Estimating MUP Conditional Probabilities

A muscle under test can be characterized by combining the set of characterizations of the MUPs detected from it. Consider that the clinical state of a muscle can be assigned to one of K specific categories with labels $\{y_1, \cdots, y_k, \cdots, y_K\}$ (i.e. myopathic, normal, or neuropathic). In this work, the characterization of a MUP is defined by a set of K MUP conditional probabilities, one for each category. Each MUP conditional probability $P(y_k \mid MUP)$ measures the probability of category y_k given the detected MUP. Therefore a MUP characterization can be expressed as a set of conditional probabilities $\{P(y_1 \mid MUP_i) \ldots P(y_K \mid MUP_i)\}$.

5.1.1 Compound Rule Conditional Probability for PD

The WOE that PD produces needs to be expressed as a conditional probability to obtain a muscle characterization. This conditional probability is called the Compound Rule Conditional Probability (CRCP). The CRCP [52, 58, 59] is an estimate of the conditional probability that MUP_i was detected

in a muscle of category y_k given the occurrence of the feature values associated with the compound rule x_k^* and is expressed as

$$P_k(MUP_i = y_k \mid MUP_i = x_k^*) = \frac{1}{\left[\left(2^{-WOE}\right)\cdot\left(\frac{1-P_0(y_k)}{P_0(y_k)}\right)\right]+1} \qquad (5.1)$$

where:

$MUP_i = y_k$, MUP_i was detected in a muscle of category y_k (i.e. a normal or neuropathic muscle).

x_k^* Compound rule associated with category y_k.

$MUP = x_k^*$, MUP feature values that match the compound rule x_k^* (i.e. a set of quantized feature values)

WOE Weight of Evidence of the compound rule x_k^* supporting or refuting category y_k.

$P_0(y_k)$, prior probability of category y_k.

Equation 5.1 is derived in Appendix A.

The conditional probability of category y_k given detected MUP_i can be estimated using the following normalization so that the conditional probabilities across all categories given MUP_i sum to one:

$$P(y_k \mid MUP_i) \cong \frac{P_k(MUP_i = y_k \mid MUP_i = x_k^*)}{\sum_{j=1}^{K} P_j(MUP_i = y_j \mid MUP_i = x_j^*)} \qquad (5.2)$$

Table 5.1 provides example component rules for estimating $P(y_k \mid MUP_i)$ assuming that the features amplitude, duration, area, thickness, and size index are used for characterizing a MUP using two categories: normal or neuropathic. In this example, each of the feature values are quantized into one of five intervals called very low, low, medium, high, and very high. Also, a prior probability of 0.5 is assumed for each category. A MUP with very high amplitude, low duration, very high area, low thickness, and high size index is compared against the rules that were created by the training data. Two component rules are found for the normal category and two component rules are found for the neuropathic category and are shown in Table 5.1. The WOE for a component rule is calculated by counting the number of instances in the training data that match the rule's category and counting the number of MUPs in the training data that do not match the rule's category to empirically determine the probabilities of belonging to that category and not belonging to that category. For instance, the first rule in Table 5.1 has 200 MUPs with very high amplitude and low thickness labeled neuropathic and 100 MUPs with very high amplitude and low thickness labeled normal in training data that has 1000 MUPs in total for each category and so using equation 4.4:

$$WOE = \log_2\left(\frac{P((amplitude = veryhigh)\ \&\ (thickness = low) \mid Category = Normal)}{P((amplitude = veryhigh)\ \&\ (thickness = low) \mid Category \neq Normal)}\right) = \log_2\left(\frac{100/1000}{200/1000}\right) = -1$$

The compound rule that refutes the normal category has a WOE of -3.7 (found by summing the two component rules for normal). Note that the WOE refutes normal because it is negative. Using equation 5.1 the CRCP for normal for this MUP is 0.0714. The compound rule that supports the neuropathic category has a WOE of 1.61 with a CRCP of 0.75325. Using Equation 5.2 to normalize the Compound Rule Conditional Probabilities to add to 1 results in a MUP Characterization of P(Normal | MUP) = 0.08 and P(Neuropathic | MUP) = 0.92.

Further details of the Pattern Discovery based method for MUP characterization [52, 58, 59] are available in Chapter 4.

Table 5.1 Example Rules for Calculating a MUP Characterization

Component Rule	Category	WOE
Very high amplitude and low thickness	Normal	-1
Low duration and very high area	Normal	-2.7
Very high amplitude and very high area	Neuropathic	2.61
Low thickness and high size index	Neuropathic	-1

5.1.2 LDA Based MUP Characterization

Probability estimates of MUP features under test were found by assuming that the feature values were Gaussian distributed [1, 11]. LDA Minimum Euclidean Distance (MED) assumes that each category has equal covariance and calculates the conditional probability of each category given a test MUP_i ($P(y_k \mid MUP_i)$) by finding the Euclidean distance of the discriminant coordinates of the test MUP to the centroid of the set of discriminant coordinates in the training data of each category and then calculating the probability assuming Gaussian distribution.

5.2 Muscle Characterization Methods

Probabilistic muscle characterization was compared to conventional muscle characterization methods.

5.2.1 Probabilistic Muscle Characterization

Two probabilistic methods for combining MUP characterizations were evaluated in this chapter: arithmetic mean and Bayes' rule.

Arithmetic Mean Muscle Characterization

The arithmetic mean characterization measure calculated for a muscle category was done by taking the arithmetic mean of the set of MUP characterizations of that category and is called *Arithmetic Mean Muscle Characterization* (AMC) and is calculated by

$$AMC(y_k) = \frac{\sum_{i=1}^{N} P(y_k \mid MUP_i)}{N} \qquad (5.3)$$

where for Equations 5.3 & 5.4:

y_k	muscle category, i.e. myopathic, normal or neuropathic.
$P(y_k \mid MUP_i)$	the MUP characterization value for *category* y_k of the i^{th} MUP detected from the muscle under test.
N	Number of MUPs detected from the muscle under test.

Bayes' Rule Muscle Characterization

In previous work Pfeiffer used Bayes' rule to combine MUP characterizations into a muscle characterization that consisted of a conditional probability that a muscle belonged to a particular category (i.e., neuropathic vs. myopathic) given the set of MUP characterizations [1]. A Bayes' rule muscle characterization measure calculated for a muscle category is done using Bayes' rule for multiple pieces of evidence, where each piece of evidence is a MUP characterization of that category and assuming each MUP characterization is conditionally independent of each other [1, 19, 52]. This is called *Bayes'-rule Muscle Characterization* (BMC) and is calculated by

$$BMC(y_k) = P(MUS = y_k \mid MUP_1, MUP_2, \ldots MUP_n) = \frac{\prod_{i=1}^{N} P(y_k \mid MUP_i)}{\sum_{j=1}^{K}\left[\prod_{i=1}^{N} P(y_j \mid MUP_i)\right]} \qquad (5.4)$$

Equation 5.4 assumes that the prior probabilities used are the same for each category to obtain an unbiased muscle characterization that is based solely on the electrophysiological evidence provided by the detected MUPs. The author believes that each MUP is adding independent information except in those cases where more than one MUP is detected from the same MU. The derivation for Equation 5.4 is found in Appendix B and Figure 5.1 provides a graphical explanation for Equation 5.4.

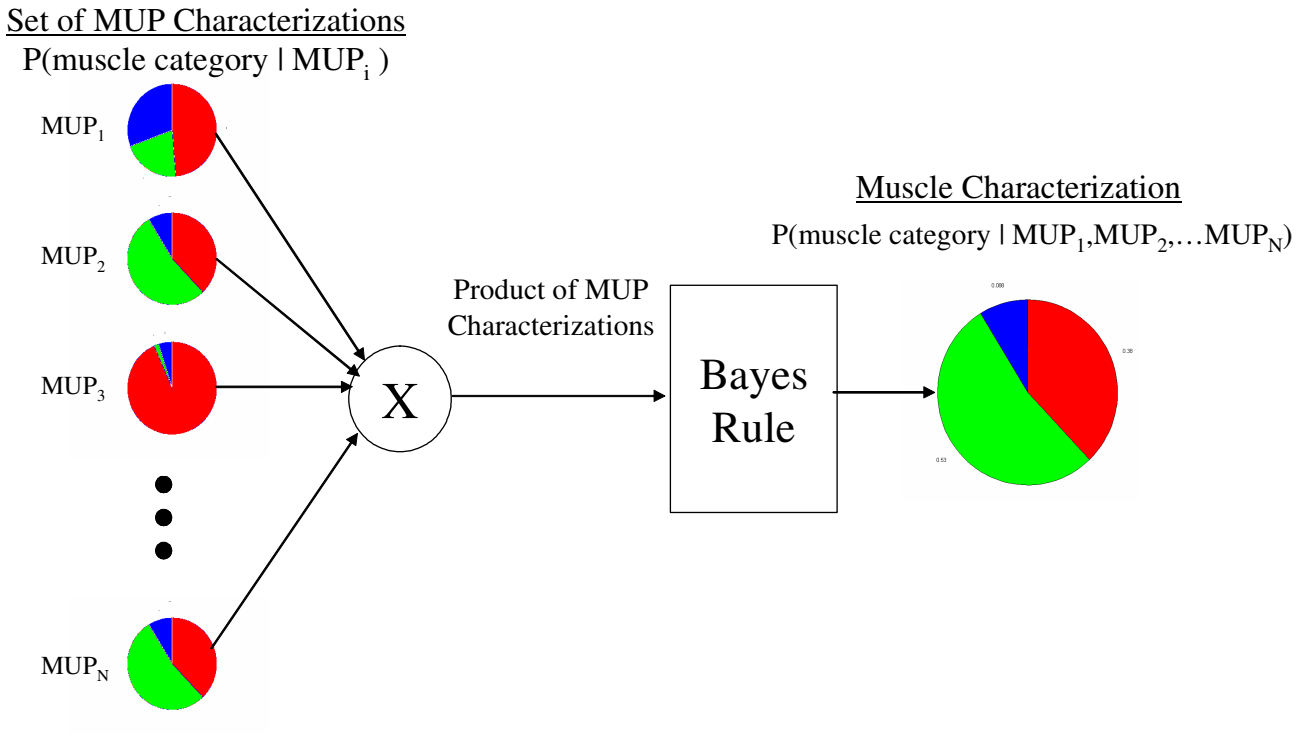


Figure 5.1 Bayes' Rule Muscle Characterization (BMC)

The pie charts on the left hand side show the characterizations of MUPs detected from the same muscle for a scenario with three muscle categories (myopathic, normal and neuropathic). Each category is represented by a colour; medium grey is myopathic, light grey is normal and black is neuropathic. The proportion of a pie is the conditional probability of a category given a detected MUP, e.g. the upper left pie shows a MUP that has approximately 50% probability of being detected

from a myopathic muscle, a 20% probability of being detected from a normal muscle and 30% probability of being detected from a neuropathic muscle. The product of the conditional probabilities of the MUPs for each category was calculated and the probability of a muscle belonging to a category given the set of MUP characterizations is determined using Equation 5.4. The larger pie chart on the right hand side shows the muscle characterization given the set of MUPs, approximately 38% myopathic, 53% normal and 9% neuropathic in this example.

5.2.2 Conventional Muscle Analysis: Means and Outlier Method

Stewart et al. developed and evaluated a technique where, given a set of MUP features, a muscle category is determined by comparing the mean value of each MUP feature, calculated across MUPs detected from a muscle under examination, to its corresponding normative threshold values [13]. The Means method calculates the mean normative threshold value using all the MUPs sampled from control muscles from which greater than 15 MUPs were detected. Threshold values for each feature are calculated using sets of 15 or more MUPs detected from the corresponding muscle of a pool of control subjects. For each set of 15 or more MUPs the mean of the feature values is calculated resulting in a mean value per feature per muscle. The overall mean and the standard deviation of the means are then calculated for each feature across the set of mean values per muscle. (Note: the overall mean and the standard deviation of the means for each feature are obtained from the set of mean values of the control muscles and not across all MUPs.) The threshold values used for the Means method for each feature are defined as the overall mean ±2 standard deviations of the means. For this work, muscles were categorized as myopathic when one or more mean feature value fell below a threshold value of the normative range and neuropathic when one or more mean feature value fell above the threshold value of the normative range.

Stalberg developed a method for categorizing a muscle by counting the number of outliers of each MUP feature from the set of the first 20 MUPs detected from a muscle under study [12]. A MUP feature value is considered an outlier if it is below or above the low and high outlier thresholds for that feature, respectively. The first 20 MUPs detected from each muscle in the pool of control subjects are used to determine outlier thresholds for each feature. Muscles with less than 20 detected MUPs are not used for establishing outlier thresholds. For each muscle and for each feature, the set of feature values of the first 20 MUPs is sorted in ascending order and the third lowest and the third highest value are added to a set of low and high outliers for that feature, respectively. The lower outlier threshold for a feature is the 5^{th} percentile of the set of low outliers and the high outlier threshold is the 95^{th} percentile of the set of high outliers. A muscle under examination for which there are three or more outlying MUP feature values of the same feature each below the lower outlier threshold is categorized myopathic. A muscle is categorized as neuropathic if three or more outlying MUP feature values of the same feature were above the higher outlier threshold.

As an additional method for muscle categorization the Means and Outlier methods are combined as in [47] and called the Combined method. If either the Means or Outlier method or both declares a myopathic/neuropathic abnormality, then the Combined method categorizes the muscle as myopathic/neuropathic.

The Means, Outlier and the Combined methods are referred to as the conventional methods.

5.3 EAS MUP Data

The Bayes' rule characterization method was used to analyze the EAS MUP data described in Section 3.3.1 and compared to the performance of the conventional methods. PD and LDA MED were used to provide MUP characterizations for the Bayes' rule method. BMC-PD refers to using Bayes' rule

using PD based MUP characterizations and BMC-LDA refers to using Bayes' rule using LDA MED based MUP characterizations.

5.3.1 Training, Testing, Features

EAS MUP Training

Performance of the BMC and the conventional methods vary depending on the data used for training. In this study, only control-reference data was used for training of the conventional methods, whereas control-reference and patient-sensitivity data was used for training of the BMC methods. Performance for all of the methods was determined across ten different trials using randomly chosen training data. The training data for a single trial of the conventional methods consisted of 116 control-reference muscles for the Means method and 100 control-reference muscles for the Outlier method. The training data for the BMC methods used all of the MUPs, from all of the 126 control-reference muscles and an equal number of MUPs sampled from the patient-sensitivity data by randomly selecting muscles and using their MUPs for training resulting in a mean number of 134 muscles being drawn from the patient-sensitivity data across the 10 trials. Table 5.2 summarizes the number of MUPs per muscle and the number of muscles used for training the various methods.

Table 5.2 Number of MUPs and Muscles used for Training: EAS MUP Data

Method	Minimum # of MUPs per muscle	# of MUPs in muscle used for training	# control-reference muscles used for training *	# patient-sensitivity muscles used for training *
Mean	15	All	116	0
Outlier	20	First 20	100	0
Bayesian	1	All	126	134

The Mean method calculates the Mean threshold value using all of the MUPs in muscles with greater than or equal to fifteen MUPs. The Outlier method uses the first twenty MUPs in muscles with

greater than or equal to twenty MUPs to calculate the Outlier threshold. The BMC methods uses all of the MUPs in a muscle for training regardless of the number MUPs detected.

*The fourth and fifth column of this Table refers to the number of control-reference and patient-sensitivity muscles used for training in the ten trial scenario as described in Section 5.3.1.

EAS MUP Testing

Only muscles from the control-specificity and patient-sensitivity groups with greater than twenty MUPs were used for testing. For each trial, all of the 113 muscles with 20 or more MUPs from the control-specificity data were used to measure specificity and all of the muscles that did not have any MUPs selected for training from the patient-sensitivity group, mean of 57 muscles across the 10 trials, were used to measure the sensitivity. For each trial, the same set of test muscles was used across the conventional and BMC methods to ensure fair comparison. The number of true negatives and true positives from each test were accumulated to arrive at the overall specificity and sensitivity results for each of the methods.

Values for the following MUP features were input to the different methods: amplitude, duration, area, number of phases, number of turns and spike duration [45, 48] as well as thickness otherwise known as area-to-amplitude ratio (AAR) [35] and size index (SI) [36]. All possible combinations of features taken two, three and four at a time were selected from the eight possible features to form feature sets. Each feature set was then used for evaluating the performance of the BMC-PD, BMC-LDA and conventional methods as described above.

5.3.2 EAS MUP Results

Table 5.3 shows the mean sensitivity, specificity, accuracy, and SSD across all possible feature sets comprised of two, three or four features.

The conventional methods had lower mean accuracy and greater difference between sensitivity and specificity than the BMC methods as highlighted by the large SSD values of the conventional methods in Table 5.3. As the number of features used for characterization increased, sensitivity increased while specificity decreased for the conventional methods, e.g. for the Outlier method sensitivity increased from 33.5% to 47% and specificity decreased from 83.1% to 74.2% for two and four features per set respectively.

With the BMC methods sensitivity and specificity were steadier than the conventional methods as the number of features used for characterization increased. The sensitivity for the BMC-LDA method increased by 1.3% and specificity for the BMC-LDA method increased by 0.2% with four features in a set as compared to two features in a set as shown in Table 5.3. The BMC-LDA method had a significant improvement in accuracy compared to the BMC-PD method according to a paired t-test across the feature sets in Table 5.3 at a significance level of 0.05.

Table 5.4 shows the best five feature sets as sorted by accuracy for each method. The conventional methods favoured area, thickness, and turns as the most discriminative features, since they occur most frequently in the top five sets. The BMC–PD favoured thickness, size index, and turns and the BMC–LDA method favoured thickness, and turns as the most discriminative features because of how often they occurred in the top five feature sets. The top five features sets for the Means method had very poor sensitivity but high specificity leading to an SSD of about 20%. The combined method had the best sensitivity and accuracy of the conventional methods when looking at the five best feature sets in Table 5.4. The best sensitivity of the BMC methods was ten percent higher than the best sensitivity of the combined method. The best specificity of the combined method was four percent higher than the best specificity of the BMC methods when looking at the five best feature sets in Table 5.4. The top five features of the combined method had lower accuracy than the BMC–LDA and BMC–PD methods.

5.3.3 Robustness of Best Feature Sets

To examine the effects of different clinics using different data for training, the best performing feature set of each method was selected for further analysis by executing 100 trials with a reduced number of control and patient muscles used for training and an increased number of patient muscles used for testing. For the BMC methods, for each trial, muscles of the control-reference data set were randomly selected until approximately 75% of its MUPs were selected. An equal number of MUPs per trial were also randomly sampled from the patient-sensitivity data by random muscle selection. The mean number of muscles used for training the BMC methods across all of the trials was 95 from the control-specificity data and 101 from the patient-sensitivity data. For the conventional methods, for each trial, a subset of the control-reference muscles chosen for training the BMC methods that met the criteria as shown in the second column of Table 5.2 were selected to calculate the mean and outlier threshold values. The mean number of muscles used for training the Mean and the Outlier methods across all of the trials was 88 and 76 respectively from the control-reference data. The test data per trial was selected using muscles not used for training resulting in 113 muscles being tested in the control-specificity data and a mean of 83 muscles across the 100 trials for the patient-sensitivity data.

Table 5.5 addresses the robustness of the methods by showing how performance varied for the best feature set of each method across a larger number of trials of different randomly chosen training data that was used to determine the results in Table 5.4. The results in Table 5.5 were determined using 100 trials and approximately 75% of the training data while Table 5.4 used 10 trials and 100% of the training data. For each feature set used, the standard deviation of sensitivity and specificity across the 100 trials is reported. Standard deviation across 100 trials does not appear to be significantly different than the SD shown in Table 5.4 for the best feature sets. Table 5.5 shows an expected reduction in accuracy because of the reduced amount of training data used compared to

Table 5.4. The difference in accuracy of Table 5.5 compared to Table 5.4 is shown in brackets for each method: Mean (-0.2%), Outlier (-1%), Combined (-0.7%), BMC-PD (-1.5%), BMC-LDA (-2%).

Table 5.3 Average Performance across All Possible Feature Combinations of EAS Data

Method		# Features used for testing					
		Two		Three		Four	
		Mean	SD	Mean	SD	Mean	SD
Mean	Sens.	29.2%	4.1%	35.1%	4.1%	39.2%	3.5%
	Spec.	90.5%	2.9%	87.6%	2.8%	85.3%	2.7%
	Acc.	59.9%	2.0%	61.4%	1.7%	62.2%	1.7%
	SSD	30.6%	3.0%	26.2%	3.1%	23.1%	2.6%
Outlier	Sens.	33.5%	7.4%	41.3%	6.8%	47.0%	6.3%
	Spec.	83.1%	5.6%	78.0%	4.9%	74.2%	4.2%
	Acc.	58.3%	1.9%	59.7%	1.6%	60.6%	1.5%
	SSD	24.8%	6.3%	18.3%	5.7%	13.6%	5.1%
Combined	Sens.	41.3%	5.5%	49.6%	5.4%	55.2%	5.1%
	Spec.	80.0%	5.5%	74.4%	4.9%	70.1%	4.2%
	Acc.	60.7%	2.2%	62.0%	2.0%	62.6%	2.0%
	SSD	19.3%	5.0%	12.4%	4.7%	7.7%	4.0%
BMC-PD	Sens.	68.1%	5.3%	65.4%	4.6%	63.6%	5.2%
	Spec.	57.2%	7.0%	61.4%	4.6%	61.6%	3.7%
	Acc.	62.6%	2.6%	63.4%	1.9%	62.6%	2.4%
	SSD	7.2%	4.2%	4.6%	2.5%	4.5%	1.6%
BMC-LDA	Sens.	65.3%	3.5%	66.0%	3.3%	66.6%	2.7%
	Spec.	63.1%	4.1%	63.1%	2.8%	63.3%	2.2%
	Acc.	64.2%	2.0%	64.5%	1.9%	64.9%	1.3%
	SSD	3.8%	1.4%	3.4%	1.1%	3.4%	1.0%

BMC–PD represents Bayesian method using PD to estimate conditional MUP probabilities. BMC–LDA represents Bayesian method using LDA to estimate conditional MUP probabilities. The standard deviation (SD) is calculated across the different feature sets

Table 5.4 Five Most Accurate Feature Sets per Method: EAS MUP Data

Method	Feature Sets	Sens (%)	SD (ten trials)	Spec (%)	SD (ten trials)	Acc (%)	SD (ten trials)	SSD (%)	SD (ten trials)
Mean	area/thick/phases	45.5%	5.3%	86.7%	0.0%	66.2%	2.7%	20.5%	2.7%
	area/thick/turns	43.3%	5.6%	88.5%	0.0%	65.9%	2.8%	22.6%	2.8%
	area/thick/SI/phases	44.7%	4.4%	86.7%	0.0%	65.7%	2.2%	21.0%	2.2%
	area/thick/phases/turns	46.4%	4.9%	85.0%	0.0%	65.6%	2.5%	19.3%	2.5%
	area/thick/turns/spike_dur	44.2%	3.7%	86.7%	0.0%	65.4%	1.8%	21.3%	1.8%
Outlier	dur/area/phases/turns	53.2%	4.1%	75.2%	0.0%	64.2%	2.1%	11.0%	2.1%
	area/thick/SI/turns	57.5%	4.6%	69.9%	0.0%	63.7%	2.3%	6.2%	2.3%
	area/thick/turns	56.7%	5.0%	70.8%	0.0%	63.7%	2.5%	7.1%	2.5%
	thick/SI/turns/spike_dur	57.5%	5.1%	69.9%	0.0%	63.7%	2.6%	6.2%	2.6%
	area/thick/turns/spike_dur	57.4%	6.7%	69.9%	0.0%	63.7%	3.3%	6.2%	3.3%
Combined	dur/area/phases/turns	61.9%	4.2%	71.7%	0.0%	66.8%	2.1%	4.9%	2.1%
	area/thick/phases/turns	63.9%	4.6%	69.0%	0.0%	66.5%	2.3%	2.9%	1.7%
	area/thick/turns/spike_dur	63.9%	6.1%	69.0%	0.0%	66.5%	3.1%	3.2%	2.3%
	area/thick/turns	62.1%	5.0%	70.8%	0.0%	66.4%	2.5%	4.5%	2.3%
	thick/SI/turns/spike_dur	63.8%	3.6%	69.0%	0.0%	66.4%	1.8%	2.8%	1.5%
BMC - PD	SI/phases/spike_dur	71.2%	5.9%	63.5%	4.0%	67.4%	2.0%	4.3%	4.1%
	area/thick/phases/turns	68.1%	6.6%	66.5%	3.7%	67.3%	2.5%	3.8%	2.7%
	thick/SI/phases/turns	69.1%	3.5%	65.5%	2.1%	67.3%	1.9%	2.0%	2.0%
	area/thick/SI/turns	70.9%	5.9%	63.6%	3.2%	67.2%	3.8%	3.6%	2.8%
	thick/SI/turns	72.0%	8.6%	61.4%	3.7%	66.7%	3.3%	7.1%	2.9%
BMC - LDA	thick/phases/turns	74.1%	2.7%	66.2%	1.7%	70.2%	1.5%	4.0%	1.8%
	dur/turns	71.8%	6.4%	64.9%	1.8%	68.3%	2.8%	4.1%	2.9%
	area/thick/SI/turns	70.3%	6.8%	65.5%	1.7%	67.8%	3.4%	3.4%	2.5%
	thick/turns	67.8%	6.2%	67.8%	1.6%	67.8%	2.8%	2.8%	2.2%
	area/thick/turns	70.2%	5.0%	65.2%	1.2%	67.7%	2.2%	2.6%	2.8%

Table 5.5 Performance of the Best Feature Sets Across 100 Trials: EAS MUP Data

Method	Feature Sets	Sens (%)	SD (100 trials)	Spec (%)	SD (100 trials)	Acc (%)	SD (100 trials)	SSD (%)	SD (100 trials)
Mean	area/thick/phases	45.5%	5.3%	86.6%	2.0%	66.0%	2.2%	20.6%	3.3%
Outlier	dur/area/phases/turns	50.6%	4.5%	75.8%	2.0%	63.2%	2.0%	12.6%	2.8%
Combined	dur/area/phases/turns	60.1%	4.0%	72.1%	1.9%	66.1%	1.9%	6.1%	2.4%
BMC-PD	SI/phases/spike_dur	68.4%	5.6%	63.5%	4.5%	65.9%	2.0%	4.2%	3.2%
BMC- LDA	thick/phases/turns	71.1%	4.6%	65.3%	2.8%	68.2%	2.2%	3.5%	2.4%

5.4 LHS MUP Data

5.4.1 LHS MUP Data Training, Testing, Features

The LHS MUP data described in Section 3.2.2 was studied in this subsection. Performance for all of the methods was determined across ten different trials using randomly chosen training data. Only control muscles were selected for training the conventional methods, whereas muscles from both control and patient groups were used for training of the Probabilistic methods. Within a trial, all muscles were tested one at a time. All of the patient muscles, except the patient muscle under test, were used for training of the Probabilistic methods and the muscle under test was classified (or categorized) as being of the category with the maximum muscle characterization value. A set of control muscles were selected at random until the number of MUPs in the patient training data equaled the number of MUPs from the control muscles. The set of control muscles used for training of the Means method and the control muscles used for training the Probabilistic method were the same per trial. All muscles in the control data had 20 or greater MUPs except one muscle that had 19 MUPs. The set of control muscles used for training the Outlier method was identical to that used for training the Means and Probabilistic methods with the exception of trials that included the muscle with 19 MUPs. It had to be excluded from the Outlier method training since it requires 20 or greater MUPs. .

All muscles in the data were used for testing of the conventional and Probabilistic methods regardless of the number of MUPs detected per muscle. Values for the following QEMG-MUP features were input to each of the different methods: amplitude, duration, area, number of phases, number of turns [7]. In addition, thickness [35] and revised size index (rSI) [37] were also used as features. All possible combinations of feature sets taken two, three, four and five at a time were examined for sensitivity, specificity and accuracy to determine which combination of features used simultaneously is the best choice for a sensitive and specific muscle categorization. Each different

feature set was subjected to ten trials, where the set of training data used per trial was the same from feature set to feature set. The sensitivity across 22 muscles and the specificity across 54 muscles were recorded for each trial. Accuracy and SSD were then calculated for each trial. AMC-PD and BMC-PD refers to using the AMC and BMC methods respectively with PD based MUP characterizations. AMC-PD, BMC-PD Probabilistic muscle characterization and the conventional methods were studied using the LHS MUP data.

5.4.2 LHS MUP Data Results

Table 5.6 shows the mean sensitivity, specificity, accuracy, and SSD across all possible feature sets comprised of two, three, four or five features.

When the results across all feature sets of five were compared, the last column of Table 5.6, the AMC-PD Probabilistic method had the highest mean accuracy of 84.5% followed by the Means and BMC-PD method whose mean accuracies were 80.3% and 78.8% respectively. The Outlier and Combined methods had lower mean accuracy than the Means and Probabilistic methods regardless of the number of features in a feature set. As the number of features used for categorization increased, sensitivity increased while specificity decreased for the conventional methods. For example, the Outlier method sensitivity increased from 56.8% to 75.9% and specificity decreased from 79.2%% to 63.2% for two and five features per set respectively. In general, the conventional methods had a greater difference between sensitivity and specificity as highlighted by their larger SSD values in Table 5.6 compared to the Probabilistic methods. With the Probabilistic methods, sensitivity and specificity both increased and demonstrated greater consistency, unlike the conventional methods, as the number of features used for categorization increased, e.g. for the AMC-PD method sensitivity increased by 3.3% while specificity increased by 5.3% when comparing results of two features per set to five features per set.

Table 5.7 shows the best five feature sets as sorted by accuracy for each method. The AMC-PD method had the feature set with the highest accuracy of 86.3% with 1.1% and 3.3% higher accuracy than the BMC-PD and Means method respectively. The AMC-PD method had the feature set with the lowest SSD value of 1.8% among all of the top five feature sets of the methods. Also the AMC-PD method appears to be more robust against variations in the training data as shown by the generally lower SD of accuracy for the top five feature sets.

The accuracy of the best feature set of the AMC-PD method (86.3%) was higher than that of the AMC-LDA and BMC-LDA which were 83.9% and 82.1% respectively.

All of the feature sets were sorted according to accuracy. A paired t-test was conducted between the most accurate feature set and each subsequent feature set in the ordered list starting with the second most accurate, to find the feature sets whose accuracy was not different at a 5% level of significance. In Table 5.8 "# feat sets" refers to the number of feature sets per method that had the same accuracy at a 5% level of significance. Table 5.8 shows the number of occurrences of each individual feature in the feature sets whose accuracy did not differ significantly. Table 5.8 shows that the AMC-PD method favoured duration, area and thickness as the most discriminative features while the BMC-PD method favoured area, thickness and number of turns and the Means method favoured phases, amplitude and area based on the number occurrences in the feature sets whose accuracy did not vary significantly. The number of feature sets per method whose accuracy did not vary significantly was higher for the AMC-PD and Means methods at 20 and 21 feature sets respectively as compared to the other methods.

Data for the FDI muscle from a patient with a neuropathic disorder, selected from the data described in Section 3.2.2, was used to demonstrate the Probabilistic muscle characterization methods (see Figure 5.2). Along the left edge of Figure 5.2, the template waveform and characterization of each of the 13 MUPs detected from the FDI muscle are shown. Each MUP characterization is

represented by a pie chart where the black area is proportional to the conditional probability that the MUP was detected in a neuropathic muscle and the grey area is proportional to the conditional probability that the MUP was detected in a normal muscle. The 13 MUPs were sorted in order from highest conditional probability of being neuropathic to lowest. For the AMC-PD method the MUP characterization values per category were averaged (see Equation 5.3), resulting in a characterization measure of 82% in support of a neuropathic condition and 18% in support of a normal condition. Using equation 5.4, the BMC-PD method calculated a characterization measure of 100% in support of a neuropathic condition and 0% in support of a normal condition. In both cases this muscle is correctly categorized as neuropathic. The legend has a vertical bar showing the scale that corresponds to 1000 μV. MUP 13 is useful for establishing relative scale because it is the sole MUP in the set whose probability of being detected in a normal muscle was higher than being detected in a neuropathic muscle. MUPs 1 - 12 in Figure 5.2 have relatively large amplitudes compared to MUP 13.

Using the same muscle as in Figure 5.2, Figure 5.3 demonstrates the Means and Outlier methods. The asterisk in each plot is where the mean for each feature value for the muscle under examination fall. Figure 5.3 shows that mean values of area and thickness are greater than 2 SD above the normative training data indicating a neuropathic condition. The Table headed Outlier method in Figure 5.3 shows that area and thickness each have three or greater outliers which are both indications of neuropathic muscle. As such, both methods correctly categorize this muscle as neuropathic.

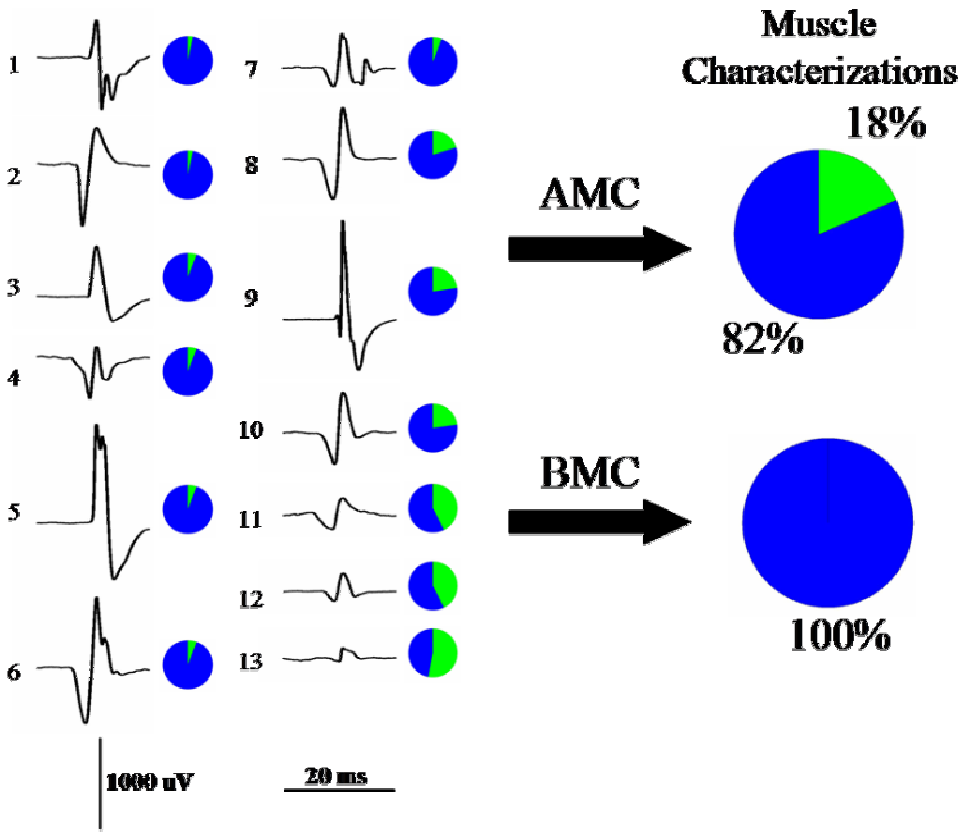


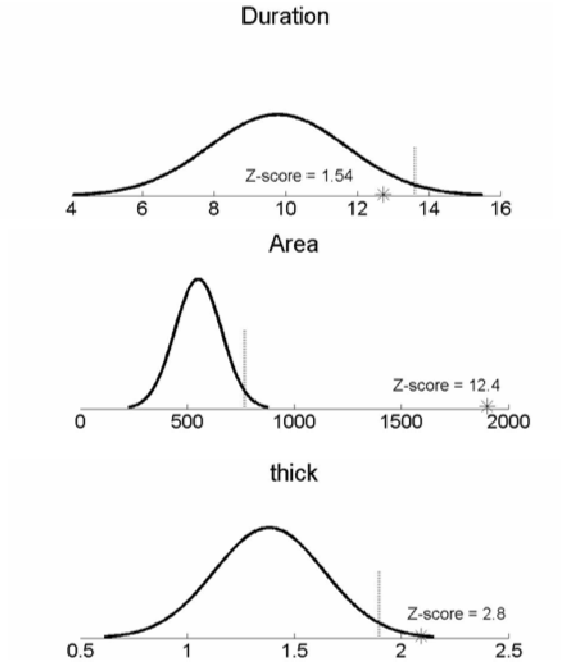
Figure 5.2 Example FDI Muscle Characterization using Probabilistic Methods

Clinician	Probabilistic Method
• qualitative	• quantitative
• prone to bias	• objective
• estimations based on examiner experience	• MUP characterization estimation based on number of occurrences in exemplary training data
• large cognitive effort needed to simultaneously consider multiple feature values across many MUPs	• computer automated

Data from an FDI muscle whose actual diagnosis is neuropathic was used to demonstrate the Probabilistic muscle characterization methods in Figure 5.2. The graphics above can also be used as an analogy of how clinicians qualitatively examine EMG signals. How a clinician performs a

qualitative electromyographic examination as compared to the probabilistic method is shown in the table in the caption under Figure 5.2. The MUP characterizations in Figure 5.2 were calculated using PD with duration, area, and thickness feature values. Area in pie charts shaded in black and grey are proportional to neuropathic and normal conditional probabilities respectively.

Mean Method



Outlier Method

	Duration	Area	Thickness
# lo Outliers	0	0	0
# hi Outliers	1	10	3

Figure 5.3 Example of Means and Outlier Analysis of an FDI Muscle

Figure 5.3 demonstrates the Means and Outlier methods using the same muscle as in Figure 5.2. The curves for the Means method were drawn assuming a Gaussian distribution of the feature values of the MUPs detected from normal muscles. The mean and standard deviation parameters used for drawing the curves were calculated using the mean and standard deviation of feature values as described in Section 5.2.2 for the Means method. The curves are meant to provide an indication of the variance of each feature value and not represent their actual distributions because the feature

values are not exactly Gaussian distributed, but assumed to be Gaussian for Means analysis. The dashed vertical line to the right of the mean of each curve is the mean plus two SD, i.e. the upper threshold of the normative training data. The asterisk in each plot is where the mean for each feature value for the muscle under examination fell. The mean value of duration across MUPs detected from the test muscle is less than 2 SD so it falls in the normative range. The means of area and thickness of the test muscle are 12.4 and 2.8 SD higher than the mean value of the normative training data for each feature respectively indicating a neuropathic condition. The Table headed Outlier method in Figure 5.3 shows that area and thickness each have three or greater outliers which are both indications of neuropathic muscle. As such, both methods correctly categorize this muscle as neuropathic.

Table 5.6 Average Performance Across All Possible Feature Sets: LHS MUP Data

		# Features used for testing							
		Two		Three		Four		Five	
Method		Mean	SD	Mean	SD	Mean	SD	Mean	SD
AMC-PD	Sens.	76.2%	7.9%	79.1%	2.6%	79.9%	2.6%	79.5%	2.6%
	Spec.	84.2%	7.5%	87.1%	3.4%	88.1%	2.8%	89.5%	2.2%
	Acc.	80.2%	7.5%	83.1%	2.3%	84.0%	1.6%	84.5%	0.9%
	SSD	4.2%	1.8%	4.2%	1.7%	4.3%	1.8%	5.1%	1.9%
BMC-PD	Sens.	76.7%	8.2%	80.2%	3.4%	80.0%	4.0%	77.4%	3.9%
	Spec.	82.9%	7.1%	83.4%	4.3%	81.2%	4.2%	80.2%	4.1%
	Acc.	79.8%	7.3%	81.8%	2.2%	80.6%	2.7%	78.8%	2.7%
	SSD	3.5%	2.0%	3.3%	1.7%	3.5%	1.5%	4.2%	1.2%
Mean	Sens.	66.4%	10.3%	73.6%	3.7%	77.0%	2.6%	79.2%	2.0%
	Spec.	89.4%	2.1%	86.2%	2.0%	83.6%	1.9%	81.3%	1.7%
	Acc.	77.9%	4.8%	79.9%	1.7%	80.3%	1.2%	80.3%	1.1%
	SSD	11.5%	5.7%	6.3%	2.4%	3.7%	1.4%	2.6%	0.6%
Outlier	Sens.	56.8%	14.3%	66.7%	9.1%	72.4%	5.9%	75.9%	4.1%
	Spec.	79.2%	4.3%	72.6%	3.9%	67.4%	3.5%	63.2%	3.0%
	Acc.	68.0%	6.4%	69.7%	4.4%	69.9%	3.3%	69.5%	2.5%
	SSD	11.4%	8.3%	5.2%	3.9%	4.3%	1.5%	6.6%	2.1%
Combined	Sens.	71.1%	12.2%	79.2%	4.9%	83.1%	2.8%	85.4%	1.9%
	Spec.	74.2%	4.5%	66.7%	3.9%	61.0%	3.3%	56.6%	2.7%
	Acc.	72.6%	5.0%	72.9%	2.7%	72.1%	2.2%	71.0%	1.8%
	SSD	5.2%	6.0%	6.8%	2.3%	11.0%	2.2%	14.4%	1.5%

Sensitivity increased and specificity decreased for each conventional method (Means, Outlier and Combined) as the number of features used for categorization increased. Sensitivity and specificity increased or remained steady for the AMC-PD Probabilistic method as the number of features used for categorization increased.

Table 5.7 Five Most Accurate Feature Sets per Method: LHS MUP Data

Method	Feature Sets	Sens (%)	Sens SD	Spec (%)	Spec SD	Acc (%)	Acc SD	SSD	SSD SD
AMC-PD	dur area thick phases turns	84.5%	2.3%	88.1%	1.6%	86.3%	1.5%	1.8%	1.3%
	dur area thick turns	83.2%	3.1%	89.3%	2.4%	86.2%	1.8%	3.2%	1.9%
	dur area thick phases	84.1%	3.2%	87.8%	2.2%	85.9%	1.8%	2.1%	1.7%
	amp dur thick SI turns	81.8%	2.1%	90.0%	1.0%	85.9%	1.0%	4.1%	1.4%
	dur area thick SI	81.8%	3.0%	90.0%	1.6%	85.9%	2.1%	4.1%	1.2%
BMC-PD	area thick turns	81.4%	3.4%	89.1%	2.5%	85.2%	2.7%	3.9%	1.3%
	area thick	81.4%	3.4%	88.7%	2.5%	85.0%	2.7%	3.7%	1.3%
	area phases turns	80.5%	2.2%	88.1%	2.0%	84.3%	1.5%	3.8%	1.5%
	area phases	81.4%	3.4%	87.2%	1.8%	84.3%	1.9%	2.9%	1.9%
	amp dur area turns	85.5%	4.7%	83.0%	3.2%	84.2%	3.2%	2.2%	1.5%
Mean	amp thick phases	79.5%	3.9%	86.5%	2.1%	83.0%	2.0%	3.5%	2.4%
	amp area thick phases	80.9%	4.2%	84.8%	2.3%	82.9%	2.1%	2.1%	2.5%
	amp dur area phases	80.0%	5.7%	85.2%	1.7%	82.6%	2.3%	3.2%	3.0%
	amp dur phases	77.7%	5.4%	87.4%	1.5%	82.6%	2.2%	5.0%	3.1%
	amp area thick SI phases	80.9%	4.2%	83.9%	2.3%	82.4%	2.2%	2.0%	2.1%
Outlier	area SI turns	77.3%	6.4%	78.0%	3.5%	77.6%	3.3%	3.4%	1.7%
	area SI phases	77.3%	6.4%	78.0%	3.1%	77.6%	3.6%	3.1%	1.6%
	area SI	72.7%	6.4%	82.2%	3.5%	77.5%	3.5%	5.2%	3.1%
	SI turns	73.6%	5.6%	80.4%	3.3%	77.0%	3.5%	3.6%	2.7%
	SI phases	73.2%	6.2%	80.4%	2.9%	76.8%	3.9%	3.6%	2.9%
Com-bined	area SI phases	84.5%	5.3%	73.7%	2.9%	79.1%	3.0%	5.4%	3.1%
	SI phases	79.5%	5.4%	77.2%	2.6%	78.4%	3.2%	2.4%	1.5%
	area SI	80.0%	5.3%	76.5%	3.3%	78.2%	3.0%	2.6%	2.5%
	area phases	76.4%	5.2%	79.1%	2.6%	77.7%	2.3%	2.6%	2.5%
	amp SI phases	85.5%	6.7%	69.6%	3.0%	77.5%	4.6%	7.9%	2.4%

Table 5.8 Number of Occurrences per Feature in Best Feature Sets: LHS MUP Data

	AMC-PD	BMC-PD	MEAN	Outlier	Combined
amp	8	5	14	1	3
dur	19	5	10	0	0
area	18	10	14	5	5
thick	13	7	10	0	0
SI	8	5	10	8	8
phases	6	4	21	5	5
turns	9	7	0	4	2
# feat sets	**20**	**15**	**21**	**8**	**9**

The number of occurrences of each individual feature in the feature sets whose accuracy did not differ significantly from the most accurate feature set is shown per each method. The term "# feat sets" refers to the number of feature sets per method that had the same accuracy at a 5% level of significance.

5.5 Rigs MUP Data Results

The MUP data described in Section 3.2.3 was studied in this section. The AMC characterization method was used to analyze the Rigs MUP data and compared to the performance of the conventional methods. PD was used to provide MUP characterizations for the AMC rule method. AMC-PD refers to using the AMC method using PD based MUP characterizations.

The same process of training and testing and selecting muscles per trial described in Section 5.4 (LHS MUP data) was used for the Rigs MUP data.

Table 5.9 shows the mean per-category accuracy, overall accuracy, and SSD across all possible feature sets comprised of two to five features.

The conventional methods had lower mean overall accuracy and larger SSD values than the AMC-PD method as seen in Table 5.9. Across all feature sets of five, AMC-PD had 10% greater accuracy and less than half the SSD than the Mean method which had the highest accuracy of the conventional methods. As the number of features used for characterization increased, myopathic and

neuropathic accuracy increased while normal accuracy decreased for the conventional methods, e.g. for the Mean method myopathic accuracy increased from 62% to 75% and normal accuracy decreased from 89% to 73% for two and five features respectively. With AMC-PD average myopathic accuracy increased from 77% to 84% while average neuropathic accuracy remained at 95% across all feature sets of two to all feature sets of five. Average normal accuracy increased for the AMC-PD method as the number of features increased unlike the conventional methods.

Table 5.10 shows the best five feature sets as sorted by accuracy for each method. The best feature set of the AMC-PD method had 7% higher accuracy than the best feature set of the Mean method. The Mean method achieved significantly lower myopathic accuracy than the AMC-PD resulting in high values of SSD of the Mean method compared to the AMC-PD method whose SSD ranged from 2% to 4%. All of the feature sets were ranked according to accuracy. A paired t-test was conducted between the most accurate feature set and each subsequent feature sets in the ordered list starting with the second most accurate, to find the feature sets whose accuracy was not different at a 5% level of significance. The Mean methods favoured SI, area, thickness, and turns as the most discriminative features, since they occur most frequently in the feature sets whose accuracy did not differ significantly. The AMC-PD method favoured thickness, turns and duration as the most discriminative features of the feature sets not differing significantly in accuracy.

Table 5.9 Average Performance Across All Possible Feature Sets: Rigs MUP Data

		# Features used for testing							
		Two		Three		Four		Five	
Method		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Mean	myo	62%	12%	69%	8%	73%	7%	75%	5%
	norm	89%	9%	84%	10%	78%	10%	73%	9%
	neur	93%	12%	97%	2%	98%	1%	98%	1%
	overall	81%	7%	84%	3%	83%	3%	82%	3%
	SSD	16%	5%	13%	3%	13%	2%	13%	2%
Outlier	myo	42%	8%	51%	5%	55%	4%	57%	3%
	norm	74%	5%	69%	6%	66%	6%	63%	5%
	neur	87%	12%	92%	3%	93%	1%	93%	1%
	overall	68%	6%	70%	3%	71%	2%	71%	2%
	SSD	20%	4%	17%	2%	17%	2%	16%	1%
Combined	myo	60%	11%	63%	7%	65%	5%	65%	5%
	norm	69%	7%	61%	8%	56%	7%	51%	6%
	neur	93%	14%	97%	3%	98%	2%	98%	1%
	overall	74%	8%	74%	3%	73%	2%	71%	2%
	SSD	16%	2%	18%	2%	19%	2%	20%	2%
AMC-PD	myo	77%	14%	83%	7%	85%	5%	84%	5%
	norm	87%	8%	90%	8%	93%	7%	95%	5%
	neur	95%	3%	95%	4%	95%	3%	95%	3%
	overall	86%	6%	89%	4%	91%	3%	92%	3%
	SSD	10%	5%	7%	3%	6%	2%	6%	2%

Table 5.10 Five Most Accurate Feature Sets per Method: Rigs MUP Data

Method	Feature Sets	myo	SD (ten trials)	norm	SD (ten trials)	neur	SD (ten trials)	overall	SD (ten trials)	SSD	SD (ten trials)
Mean	area SI phases	70%	5%	99%	3%	100%	0%	90%	2%	14%	2%
	SI turns	74%	6%	94%	5%	99%	2%	89%	3%	11%	2%
	SI phases	68%	7%	100%	0%	98%	2%	89%	2%	15%	3%
	dur area SI turns	82%	5%	83%	7%	100%	0%	88%	2%	9%	2%
	area SI turns	72%	3%	93%	7%	100%	0%	88%	3%	12%	1%
Outlier	amp dur area turns	58%	8%	80%	0%	93%	1%	77%	2%	15%	3%
	amp area turns	57%	9%	80%	0%	93%	1%	76%	3%	15%	4%
	area turns	56%	7%	80%	0%	93%	1%	76%	2%	15%	3%
	amp dur turns	57%	9%	80%	0%	90%	2%	76%	3%	14%	4%
	dur area turns	54%	8%	80%	0%	93%	1%	76%	3%	16%	4%
Combined	thick SI	74%	6%	70%	0%	100%	0%	81%	2%	14%	1%
	dur area	60%	6%	79%	3%	100%	0%	80%	3%	17%	2%
	dur area turns	65%	8%	73%	7%	100%	0%	79%	4%	15%	3%
	dur SI	66%	7%	70%	0%	100%	0%	79%	2%	15%	2%
	dur area SI	66%	7%	69%	3%	100%	0%	78%	3%	15%	2%
AMC-PD	amp dur thick phases turns	95%	0%	100%	0%	96%	0%	97%	0%	2%	0%
	dur thick turns	95%	0%	100%	0%	96%	0%	97%	0%	2%	0%
	dur area thick phases turns	92%	3%	100%	0%	98%	3%	97%	1%	4%	2%
	amp thick turns	95%	0%	94%	7%	100%	0%	96%	2%	4%	2%
	dur thick phases turns	94%	2%	100%	0%	95%	2%	96%	1%	3%	1%

5.6 University of Ljubljana Biceps MUP Data

The performance of the conventional method was compared against the AMC-PD method on the Ljubljana biceps MUP data described in Section 3.3.2. This section will also examine the effect of using different methods for establishing the normative limits for the Mean method. Previous work in this section examined the mean across pooled muscle data (the set of mean muscle values per feature) ± 2 SD as was discussed in Section 5.2.2. This section also discusses the effect of using broader limits for the Mean method by taking the 5^{th} and 95^{th} and then 2.5^{th} and 97.5^{th} percentiles of the pooled muscle means because Podnar et al. examined the effect of broader limits of the Mean method on sensitivity using the Ljubljana biceps MUP data set studied in this section [51]. The same set of features were used as defined for the EAS MUP data in Section 5.4.1 except spike duration was not used for testing of the biceps MUP data because it was not available for many of the muscles.

5.6.1 Ljubljana Biceps Data Results

The conventional methods had lower mean accuracy and greater difference between sensitivity and specificity than the AMC-PD method as highlighted by the large SSD values of the conventional methods in Table 5.11. As the number of features used for characterization increased, sensitivity increased while specificity decreased for the conventional methods, e.g. for the Mean 2.5^{th} to 97.5^{th} percentile method sensitivity increased from 27.1.5% to 58.8% and specificity decreased from 93.9% to 84.6% for one and five features per set respectively. The AMC-PD mean sensitivity and specificity both increased as the number of features per set increased to 83.3% mean sensitivity, 89.5% mean specificity and mean accuracy of 86.4% across all sets of five features.

Table 5.12 shows the best five feature sets per method as sorted by accuracy. The most accurate feature set of the AMC-PD method was 10% more accurate than the best feature set of the Combined method which had the highest accuracy of the conventional methods. All of the feature sets were

sorted according to accuracy. A paired t-test was conducted between the most accurate feature set and each subsequent feature set in the ordered list starting with the second most accurate, to find the feature sets whose accuracy was not different at a 5% level of significance. AMC-PD favoured thickness, area and duration as the most discriminative features. The combined method had the feature set with the highest accuracy of the conventional methods. The combined method favoured thickness, area and SI as the most discriminative feature sets.

Table 5.11 Average Performance Across All Possible Feature Sets: Ljubljana Biceps MUP Data

		# Features used for testing									
		One		Two		Three		Four		Five	
Method		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Mean +/- 2SD	Sens.	18.6%	20.9%	31.5%	19.6%	40.2%	16.6%	46.1%	13.6%	50.2%	11.1%
	Spec.	96.5%	1.1%	93.9%	1.7%	91.8%	2.1%	90.0%	2.1%	88.2%	2.0%
	Acc.	57.6%	10.7%	62.7%	10.0%	66.0%	8.3%	68.0%	6.6%	69.2%	5.3%
	SSD	38.9%	10.2%	31.2%	9.7%	25.8%	8.4%	21.9%	7.1%	19.0%	5.9%
Mean 5th to 95th pctle	Sens.	33.8%	24.8%	51.4%	17.5%	60.6%	11.3%	65.4%	7.8%	68.1%	6.0%
	Spec.	89.1%	2.4%	81.9%	2.3%	77.1%	2.3%	73.9%	2.2%	71.6%	1.9%
	Acc.	61.5%	11.5%	66.7%	8.4%	68.9%	5.8%	69.6%	4.2%	69.8%	3.4%
	SSD	27.6%	13.3%	15.3%	9.2%	8.5%	5.6%	4.9%	3.3%	3.2%	1.9%
Mean 2.5th to 97.5th pctle	Sens.	27.1%	23.0%	41.9%	17.7%	50.0%	13.6%	55.0%	11.6%	58.8%	10.0%
	Spec.	93.9%	3.0%	90.3%	2.6%	88.0%	2.6%	86.1%	2.5%	84.6%	2.1%
	Acc.	60.5%	10.2%	66.1%	8.0%	69.0%	6.2%	70.6%	5.3%	71.7%	4.6%
	SSD	33.4%	12.9%	24.2%	9.8%	19.0%	7.6%	15.6%	6.5%	12.9%	5.6%
OUTLIER	Sens.	28.6%	26.5%	44.4%	22.7%	53.5%	19.0%	59.2%	15.6%	63.0%	11.4%
	Spec.	89.3%	3.9%	82.3%	3.6%	77.5%	3.6%	73.9%	3.5%	71.1%	3.0%
	Acc.	58.9%	11.9%	63.4%	10.4%	65.5%	8.7%	66.6%	7.0%	67.0%	5.0%
	SSD	30.4%	14.7%	19.1%	12.3%	12.7%	9.7%	8.9%	7.4%	6.0%	4.9%
Combined	Sens.	32.8%	28.8%	50.4%	22.7%	59.8%	17.6%	65.1%	13.8%	68.2%	10.1%
	Spec.	86.6%	4.2%	78.0%	3.5%	72.3%	3.4%	68.1%	3.4%	65.0%	3.0%
	Acc.	59.7%	12.8%	64.2%	10.7%	66.0%	8.7%	66.6%	7.0%	66.6%	5.4%
	SSD	26.9%	16.1%	14.4%	11.5%	8.4%	7.2%	5.8%	4.7%	4.7%	3.2%
AMC-PD	Sens.	72.8%	12.0%	79.4%	6.1%	80.3%	4.8%	80.8%	4.2%	83.3%	4.5%
	Spec.	70.3%	17.0%	79.0%	11.6%	85.4%	8.1%	89.5%	4.7%	89.5%	4.0%
	Acc.	71.5%	13.7%	79.2%	8.4%	82.8%	5.7%	85.2%	3.7%	86.4%	2.6%
	SSD	3.5%	4.2%	3.0%	2.6%	3.7%	2.3%	4.6%	2.0%	4.1%	1.9%

Table 5.12 Five Most Accurate Feature Sets per Method: Ljubljana Biceps MUP Data

Method	Feature Sets	Sens (%)	Sens SD	Spec (%)	Spec SD	Acc (%)	Acc SD	SSD	SSD SD
Mean +/- 2SD	dur area thick SI	59%	3%	94%	0%	77%	2%	17%	2%
	area thick SI	59%	3%	94%	0%	77%	2%	17%	2%
	dur thick SI	59%	3%	94%	0%	77%	2%	17%	2%
	thick SI	59%	3%	94%	0%	77%	2%	17%	2%
	thick	56%	3%	97%	0%	76%	2%	20%	2%
MEAN 5th to 95th pctle	thick SI	72%	4%	85%	0%	78%	2%	6%	2%
	area thick SI	73%	5%	82%	1%	77%	3%	4%	2%
	thick SI turns	72%	4%	82%	0%	77%	2%	5%	2%
	area thick	71%	7%	82%	1%	76%	3%	5%	3%
	area thick SI turns	73%	5%	78%	1%	76%	3%	3%	1%
MEAN 2.5th to 97.5th pctle	dur area thick SI turns	66%	4%	88%	0%	77%	2%	11%	2%
	area thick SI turns	66%	4%	88%	0%	77%	2%	11%	2%
	dur area thick turns	66%	4%	88%	0%	77%	2%	11%	2%
	dur area thick SI	66%	4%	88%	0%	77%	2%	11%	2%
	area thick SI	66%	4%	88%	0%	77%	2%	11%	2%
OUTLIER	thick	69%	6%	88%	0%	79%	3%	9%	3%
	thick turns	69%	6%	85%	0%	77%	3%	8%	3%
	area thick SI	72%	4%	82%	0%	77%	2%	5%	2%
	thick SI	72%	4%	82%	0%	77%	2%	5%	2%
	area thick	72%	4%	82%	0%	77%	2%	5%	2%
Combined	thick	75%	4%	85%	0%	80%	2%	5%	2%
	area thick SI	79%	6%	79%	0%	79%	3%	2%	2%
	thick SI	79%	4%	79%	0%	79%	2%	2%	1%
	area thick	79%	4%	79%	0%	79%	2%	2%	1%
	thick turns	75%	4%	82%	0%	79%	2%	4%	1%
AMC-PD	dur area thick phases turns	87%	2%	94%	1%	90%	1%	3%	1%
	dur area thick phases	87%	2%	94%	1%	90%	1%	3%	1%
	dur area thick turns	87%	0%	93%	1%	90%	1%	3%	1%
	dur area thick	87%	0%	93%	1%	90%	1%	3%	1%
	thick SI turns	86%	3%	94%	1%	90%	2%	4%	2%

5.7 Discussion of Muscle Characterizations

MUPs with specific feature values (i.e. large area, large amplitude, many turns) can in principle be detected in a myopathic, normal or neuropathic muscle, however, with different probabilities. In general, single MUPs are nonspecific for different disease states of a muscle. Instead, it is the balance or combination of MUP conditional probabilities across a set of MUPs detected in a muscle under examination that provides the information regarding the likelihood of the muscle being myopathic, normal or neuropathic. Evaluation of a set of MUPs allows the degree to which a diagnostic concept matches the set of detected MUPs to be assessed. To be accurate, MUP-based muscle characterization must be based on multiple observations that are uncertain, taken by themselves.

5.7.1 Comparison of Probabilistic and Qualitative Methods

Probabilistic muscle characterization is analogous to standard medical practice where a clinician qualitatively combines evidence presented by detected MUPs. During qualitative EMG examinations, a clinician characterizes each MUP detected and then combines these observations into an overall muscle characterization. Probabilistic muscle characterization is based on the same reasoning as used for qualitative EMG decisions.

Probabilistic muscle characterization resembles the methods clinicians use to qualitatively examine needle EMG signals. First, a clinician subjectively assesses the similarity of a specific MUP under examination to MUPs detected in muscles of specific categories and then implicitly forms an estimate of the probability of detecting this MUP in a muscle of a specific category. This is similar to the MUP characterizations represented by the smaller pie charts in Figure 5.2. Next the clinician combines the probability estimates of all MUPs examined to formulate an overall muscle characterization – similar to the muscle characterization represented by the larger pie chart to the right in Figure 5.2. However, a clinician may be prone to making biased decisions because they may look for MUPs that confirm a pre-conceived expectation or assign lesser importance to MUPs that contradict an expectation. As well, there is a possibility of quickly jumping to an incorrect conclusion based on the observation of a single MUP feature, (e.g. it is common to associate increased MUP amplitude with a neuropathic condition). Probabilistic muscle characterization is a quantitative method that uses unbiased MUP characterizations that are estimated by simultaneously considering multiple MUP features and based on numbers of occurrences in exemplary training data. A summary of the differences between clinicians using qualitative analysis and Probabilistic methods is provided in the caption of Figure 5.2.

The author believes that skilled electromyographers can achieve similar levels of accuracy as the Probabilistic methods during a qualitative examination implying that they can both correctly estimate

the MUP conditional probabilities for each MUP (one probability per MUP per category) and can successfully combine these probabilities into a single decision. However, it is expected that the speed and accuracy with which electromyographers master both of these difficult to acquire skills can be increased through the feedback provided by quantitative MUP and muscle characterization. Taking into account multiple features simultaneously is a cognitive burden for the electromyographer using qualitative methods whereas it is an inherent aspect of Probabilistic muscle characterization.

5.7.2 Specific MUP Data Sets

EAS MUP Data

As the number of features used for characterization increased, sensitivity increased while specificity decreased for the conventional methods and remained steady for the BMC-PD method.

In previous work done by Podnar, area and turns individually had high sensitivity and duration had high specificity with the EAS data [47]. When using area, duration and turns simultaneously, BMC-LDA method had an almost three percent improvement in accuracy over the combined method as shown in Table 5.13. The Means method had very poor sensitivity and very high specificity. Table 5.13 also shows that this work achieved similar results to those of Podnar [47] in implementing conventional analysis.

Regardless of the method, thickness and turns appeared often in the top five feature sets as determined by accuracy. This work agrees with findings reported by Nandedkar et al. for limb muscles [35] suggesting that thickness improves accuracy more than area or duration alone but disagrees with previous analysis [47, 60] done on the EAS MUP data described in Section 3.3.1 that did not find thickness to be more discriminative than area and duration. Previous work using this EAS MUP data did not find thickness to be discriminative because not all combinations of feature sets were tested in [47, 60].

The BMC–PD method was not as accurate as the BMC–LDA method in the study of the EAS MUP data. As shown in Chapter 4, the PD method was marginally less accurate in characterizing MUPs than the LDA method. However, the PD method for MUP characterization can provide useful information in some cases when used in combination with the BMC–LDA method. An event is significant when it occurs more often than expected assuming random occurrence. The significant events can be used by the PD method to determine sets of feature values that contributed most to the characterization of individual MUPs [61] and can be useful in explaining the basis of a characterization of a MUP (i.e. transparency). Although transparency was not the focus of the muscle characterization study conducted in Chapter 5, it is important to note that the transparent characterization of a MUP can be valuable in situations where the characterization of a MUP contradicts a clinician's initial judgment. In addition, the BMC-PD method is useful for characterizing nominal valued features – an ability the LDA method does not have.

LHS MUP Data

The AMC-PD method had the highest accuracy across all of the methods evaluated at 86.2%. As the number of features used for characterization increased, sensitivity increased while specificity decreased for the conventional methods and remained steady for the AMC-PD method. The AMC-LDA method had slightly lower accuracy than the AMC-PD method. AMC-PD method favoured area, duration and thickness as the most discriminative features.

Rigs MUP Data

The AMC-PD method had the highest accuracy across all of the methods evaluated at 97%. As the number of features used for characterization increased, myopathic and neuropathic accuracy increased while normal accuracy decreased for the conventional methods and all three categories remained steady or increased for the AMC-PD method. The AMC-PD method favoured duration, thickness and turns as the most discriminative features.

Ljubljana Biceps MUP Data

The AMC-PD method at 90% had the highest accuracy across all of the methods evaluated. As the number of features used for characterization increased, sensitivity increased while specificity decreased for the conventional methods. AMC-PD method favoured area, duration and thickness as the most discriminative features.

5.7.3 Comparison Across MUP Data Sets

As discussed in Section 3.4 the EAS MUP data set does not have a great deal of separation between its categories while the other data sets have a reasonable separation distance. The accuracy of characterizing the EAS MUP data was the lowest and the Rigs MUP data accuracy was the highest across all of the MUP data sets consistent with the separation distance of the distributions represented by their thickness values as shown in Table 3.1.

The BMC–LDA method had better accuracy with lower variance across different feature sets for the EAS MUP data than the BMC-PD method. This suggests that the BMC–LDA method may provide better estimates of the MUP conditional probabilities than the BMC–PD method for closely separated data. Even though LDA works best for continuous feature values, it was successful in using information provided by integer valued features such as number of phases and number of turns for the EAS MUP data.

When there is a reasonable spread in distance between categories as is the case for the LHS MUP data, the AMC-PD or BMC-PD methods provided higher accuracy than the AMC-LDA or BMC-LDA methods. Also AMC-PD based characterization has higher accuracy than BMC-PD methods for MUP data sets with reasonable distance of separation.

AMC-PD performed consistently while the conventional methods performance was not consistent across the MUP data sets. The Mean method did not perform well using either the EAS or Ljubljana biceps MUP data. For the EAS MUP data the Mean method had very poor sensitivity, i.e.

39% average sensitivity across all sets of four features shown in Table 5.3 and a sensitivity of 45.5% with the best feature set as shown in Table 5.4. For Ljubljana biceps MUP data the Mean method (using ± 2SD normative limits) had poor sensitivity of 59% with the best feature set as shown in Table 5.12. Table 5.12 shows that the sensitivity of the Mean method improves when using wider normative limits (5^{th} to 95^{th} percentile), however, the improved sensitivity comes at the cost of reduced specificity. Performance of the Mean method did improve for the LHS and Rigs MUP data; however, its performance did not exceed the performance of the AMC-PD method. The performance of the outlier and the combined conventional methods also had inconsistent performance across the MUP data sets. The probabilistic methods performed better than conventional methods across all of the MUP data sets. Also sensitivity, specificity and or per category accuracy remained high for the AMC-PD and BMC-PD methods regardless of the number of features used for characterization.

Thickness appeared in all and duration in all but one MUP data set as discriminative features. This suggests that duration and thickness should at the very least be included as features in the development of a CDSS regardless of the type of muscle or the suspected disease process affecting a muscle under examination.

5.7.4 Advantages of the AMC-PD Method

Pattern Discovery used three intervals to quantize feature values for the LHS and Ljubljana biceps MUP data and five intervals to quantize feature values for the EAS and Rigs MUP data. PD performed well relative to the other MUP characterization methods examined. Using a small number of intervals helps to simplify the visual patterns that explain the results leading to diagrams that are easily recognized and understood by clinicians (see Chapter 4). Using a small number of intervals, e.g. three, results in a wider range of feature values that are considered small, medium or large. This

suggests that for AMC-PD muscle characterization high levels of precision in placing the onset and end markers that define the durations of MUPs is not required to ultimately achieve a high level of muscle categorization accuracy.

The results in Table 5.5 provide some indication that the Probabilistic methods are robust to changing the size of the training data and varying the composition of the training data.

A better balance between sensitivity and specificity was obtained by the Probabilistic methods (as shown by the lower SSD values) as compared to the conventional methods. This was most likely a result of using pattern classification techniques for estimating conditional probabilities of the MUPs. Training data from both categories are used so that more information is used for the estimation of MUP conditional probabilities.

Probabilistic muscle characterization is as accurate, or more accurate, than the conventional quantitative methods. Probabilistic muscle characterization facilitates the determination of "possible" (likely), "probable" (more likely), or "definite" (most likely) levels of pathology [47] whereas the conventional methods are based on hypothesis testing and the number of criteria present which is a dichotomous "normal" or "abnormal" decision. One can see from Figure 5.2 that the clinician using AMC-PD could define numerical intervals that are based on a continuous scale corresponding to "possible", "probable", or "definite" levels of pathology. AMC-PD is based on quantitative data and can be directly used to support clinical decisions related to initial diagnosis as well as treatment and management over time. A system based on Probabilistic characterization can also formally combine other relevant clinical information to the case at hand such as the characterization of other muscles, and non-electrophysiological measurements, symptoms, and or test results.

The sensitivity and specificity of conventional analysis varied considerably as the number of features used for characterization increased. The results showed the dramatic increase in sensitivity as the number of features used for the conventional methods increased accompanied by a dramatic

decrease in specificity. However, the Probabilistic methods did not vary as the number of features used for characterization changed. This property allows more flexibility in the choice of feature sets when implementing a system based on Probabilistic methods.

The conventional Means and Outlier method required a minimum of twenty detected MUPs [47, 52] testing of a muscle. In this thesis, all of the muscles were tested regardless of the number of MUPs detected per muscle except for the EAS MUP data. Allowing all muscles to be tested regardless of the number of MUPs detected does not require that the clinician keep looking until a minimum of twenty MUPs are found – a potentially difficult task in muscles where a neuropathic process reduces the number of motor units. Note that the AMC-PD method requires a smaller number of MUPs to be detected to test a muscle.

It is expected that Probabilistic characterization can make electrophysiological examinations more objective and accurate for most electromyographers. AMC-PD provides more diagnostic information than what is provided by conventional muscle characterization techniques. In addition, Probabilistic methods are potentially able to provide measures that strongly correlate with the level of involvement of a disorder. Another measure that may strongly correlate with the level of involvement of a disorder is described in Chapter 6. The next step performed by the CDSS is to convert the scores produced by muscle characterization into well calibrated conditional probabilities of a muscle category given the set of detected MUPs and is covered in Chapter 6.

Table 5.13 Comparison of Area, Duration and Turns Performance: EAS MUP Data

	Dur/Area/Turns This Work			Dur/Area/Turns (Podnar. 2004)		
	Sens	Spec	Acc	Sens	Spec	Acc
Mean	36.4%	86.7%	61.6%	38.0%	88.0%	63.0%
Outlier	46.2%	75.2%	60.6%	43.0%	81.0%	62.0%
Combined	54.2%	73.5%	63.8%	51.0%	79.0%	65.0%
B - PD	70.0%	59.6%	64.8%	NA	NA	NA
B - LDA	66.7%	66.5%	66.6%	NA	NA	NA

Chapter 6
Calibrated Muscle Characterization

Previous chapters have focused on the accuracy muscle characterization as ratios of predicted categories to true categories in sets of test data. A muscle characterization produces a score s_k between 0 and 1 for each category y_k. Characterization accuracy can be expressed as the conditional probability of the predicted category, i.e. the category with the largest score value, given that the category of a muscle is true as shown in equation 6.1.

$$P\left(s_k > s_i \quad \forall i \neq k \quad | \, y_k = true\right) \qquad (6.1)$$

Characterization accuracy is useful information when clinicians are evaluating the performance of clinical decision support systems. Chapter 5 has shown that the AMC and BMC methods are accurate as defined by equation 6.1 but there was no evaluation of whether the scores s_k produced by the AMC and BMC methods were correlated with the level of confidence. A score s_k that is well correlated with level of confidence means that the score s_k reflects the probability that the characterization is true. Intuitively, if we consider all of the muscles characterized with a score of s_k =.8, then 80% of these muscles should actually be of the category k. Scores produced by the BMC method often saturate to either 0 or 1 as the number of MUPs used for characterization increases, see Appendix C for an explanation, so the BMC method is not considered to be well calibrated.

This chapter will discuss another muscle characterization method known as the Z-transform (ZT) method to compare the performance of its calibrated muscle characterizations to the AMC method. The chapter will then discuss a method for calibrating scores s_k and then evaluate calibrated scores produced from AMC and ZT muscle characterizations.

6.1 Z -Transform Muscle Characterization

The Z transform assumes that the conditional probabilities that comprise MUP characterizations are Gaussian distributed. The conditional probability of category c given MUP_i is called $P_{ci} = P(category = c \mid MUP_i)$. Each P_{ci} is converted into a normalized z score by subtracting the mean across all P_{ci} and dividing by the standard deviation across all P_{ci} as shown by equation 6.3. The z scores Z_{ci} are Gaussian distributed with a mean of zero and a standard deviation of 1.

$$Z_{ci} = \frac{P_{ci} - \mu}{\sigma} \qquad (6.3)$$

where

μ is the mean across all P_{ci}.

σ is the standard deviation across all P_{ci}.

The z scores (Z_{ci}) are summed across a single category and divided by the standard deviation of the sum which is the square root of the number of MUPs as shown by:

$$Z_{mc} = \sum_{i=1}^{N} Z_{ci} \Big/ \sqrt{N} \qquad (6.4)$$

where

Z_{mc} is the Z-transform of category c of muscle m across N MUPs.

N is the number of MUPs in muscle m.

The value Z_{mc} is then transformed into a probability $P(c \mid m)$ by looking up the probability value in the Gaussian cumulative density function. More information about the Z-transform can be found in [62]. In the rest of this chapter the notation y_k will be used to denote a category as has been used

earlier in this document. A muscle characterization calculated by the ZT method is a set of conditional probabilities.

Appendix D has a proof that shows that the Z transform of MUP characterizations estimated using PD have the same rankings of categories as sorted by their muscle characterization scores as compared to the AMC-PD method on a per muscle basis because the Z transform performs a linear operation on the AMC-PD characterization. This means that the Z transform method has identical sensitivity and specificity as the AMC-PD method for a given MUP data set. A pilot test showed that the ZT-PD method had identical sensitivity and specificity as the AMC-PD method using simulated MUP data.

6.2 Category Membership Probability

The characterization measures calculated by the AMC, BMC or ZT method of a muscle MUS can be thought of as scores $s_k(MUS)$ for each category y_k. Equation (6.2) is called the empirical class membership probability [63]. It is equal to the number of muscles with score s that belong to category y_k divided by the total number of muscles with score s.

$$P(y_k = true \mid s_k(MUS) = s) = \frac{(n_{ks} = y_k)}{(n_{ks} = y_k) + (n_{ks} \neq y_k)} \qquad (6.2)$$

where

$s_k(MUS)$ is the score of category y_k of muscle MUS

$n_{ks} = y_k$ is the number of muscles with score $s_k(MUS) = s$ whose true category $= y_k$

$n_{ks} \neq y_k$ is the number of muscles with score $s_k(MUS) = s$ whose true category $\neq y_k$

A plot of $P(y_k \mid s_k(MUS) = s)$ versus the score s is called a reliability diagram [64]. A muscle characterization method is well calibrated if $P(y_k \mid s_k(MUS) = s)$ converges to the score

value $s_k(MUS)$. In other words, a muscle characterization method is well calibrated if all points fall on the y=x line of the reliability diagram, i.e. the score is well correlated to the level of confidence*.

In this work, s_k denotes un-calibrated and $\hat{s}_k$ denotes calibrated scores. A reliability diagram is drawn for each category of a muscle characterization method. A function $\hat{P}(y_k \mid s_k)$ is learned for a muscle characterization method by finding a fit to the points drawn in its corresponding reliability diagrams. An un-calibrated score s_k is calibrated by changing its value to $\hat{s}_k = \hat{P}(y_k \mid s_k)$. This chapter examines the reliability diagrams of muscle characterization scores and describes the method for calibrating the scores in further detail.

6.3 Calibration Method

The reliability diagram is used for plotting the empirical category membership probability versus score for binary (2 category) classification problems. A binary classifier considers an example positive if it is a member of the category or negative otherwise. In a multiple category case a reliability diagram is drawn for each category where the category under consideration is the positive category and all the other categories are considered negative categories. Representing a multiple category problem as a set of binary problems is called the one-against-all approach [63].

---

* Reliability diagrams are often used in the literature to determine the reliability of probability of precipitation (POP) estimates provided by weather forecasters [64]. The y value of a reliability diagram for POP estimates is determined by counting the number of days precipitation actually fell on those days that the POP=x% divided by the total number of days that the POP=x%. The POP values are located on the x-axis. So, as an example, a well calibrated POP means that precipitation fell on 40% of all days that the POP value was 40%.

Mapping scores into probability estimates requires learning a mapping for each category. Category membership probabilities $P(y_k \mid s_k(MUS_i) = s)$ are learned by using muscles that actually belong to the category as positive categories and all other muscles belonging to the other categories as negative. A reliability diagram plots the following set of points for category y_k $\left(P(y_k \mid s_k(MUS_i)), s_k(MUS_i)\right)$.

A classification system is well behaved if its scores are well ranked, i.e. if $s_k(MUS_i) > s_k(MUS_j)$ then $P(y_k \mid MUS_i) > P(y_k \mid MUS_j)$. A classifier that ranks well is isotonic (monotonically non-decreasing). Therefore a mapping function that is learned also needs to be isotonic. A common way of learning an isotonic function is called the pair-adjacent-violators (PAV) algorithm [63]. This method can take a set of points $\left(P(y_k \mid s_k(MUS_i)), s_k(MUS_i)\right)$ and learn an isotonic function $\hat{P}(y_k \mid s_k)$ that fits the data using mean-squared error criteria.

The PAV algorithm learns a function $\hat{P}(y_k \mid s_k)$ as follows. The set of category membership probabilities $\{P(y_k \mid s_{ki})\}_{i=1}^{N}$ is ordered according to the ranking of their respective scores, in mathematical notation $\ldots \le s_{k(i-1)} \le s_{ki} \le s_{k(i+1)} \ldots$.

If $P(y_k \mid s_{ki}) \ge P(y_k \mid s_{k(i-1)})$ then these two points are isotonic and $\hat{P}(y_k \mid s_{ki}) = P(y_k \mid s_{ki})$. If $P(y_k \mid s_{ki}) < P(y_k \mid s_{k(i-1)})$ then the two points are known as pair adjacent violators and are both replaced by their average i.e. $\hat{P}(y_k \mid s_{ki}) = \hat{P}(y_k \mid s_{k(i-1)}) = \left(P(y_k \mid s_{ki}) + P(y_k \mid s_{k(i-1)})\right)/2$. The new estimate is checked against previous estimates to ensure that the function has remained isotonic. If not, then previous estimates are replaced by averaging until the set becomes isotonic. A Matlab function called IsoMeans supplied by Lutz Dumbgen [65] was used to implement the PAV algorithm.

A score is calibrated by assigning $\hat{s}_k = \hat{P}(y_k \mid s_k)$, i.e. looking up the y value in the learned isotonic functions that correspond to the score. At this stage in the calibration process, the set of calibrated scores $\hat{s}_k$ are unnormalized because the sum across all categories is not likely to equal 1. These calibrated and unnormalized scores are normalized by dividing each score $\hat{s}_k$ by the sum of scores across all categories, i.e. $\hat{s}_k \Big/ \sum_k \hat{s}_k$ resulting in normalized calibrated conditional probabilities.

Figure 6.1 displays an example calibration. The pie chart at the bottom shows an un-calibrated muscle characterization of 0.1 myopathic, 0.2 normal, and 0.7 neuropathic. Isotonic functions have been learned per category using training data and are labeled $\hat{P}(myo \mid score)$, $\hat{P}(norm \mid score)$ and $\hat{P}(neur \mid score)$ in Figure 6.1 for the myopathic, normal and neuropathic categories respectively. The calibrated and unnormalized score for a category is determined by finding the y value associated with the un-calibrated score in the isotonic function of that category. This results in calibrated and unnormalized scores of 0.03 myopathic, 0.1 normal and 0.85 neuropathic. These scores are normalized resulting in a calibrated muscle characterization of 3.1% myopathic, 10.2% normal and 86.7% neuropathic as shown in the top pie chart. Reliability diagrams for calibrated muscle characterizations can be drawn based on the calibrated normalized scores and performance can be checked by how close their reliability diagrams approach the y=x line.

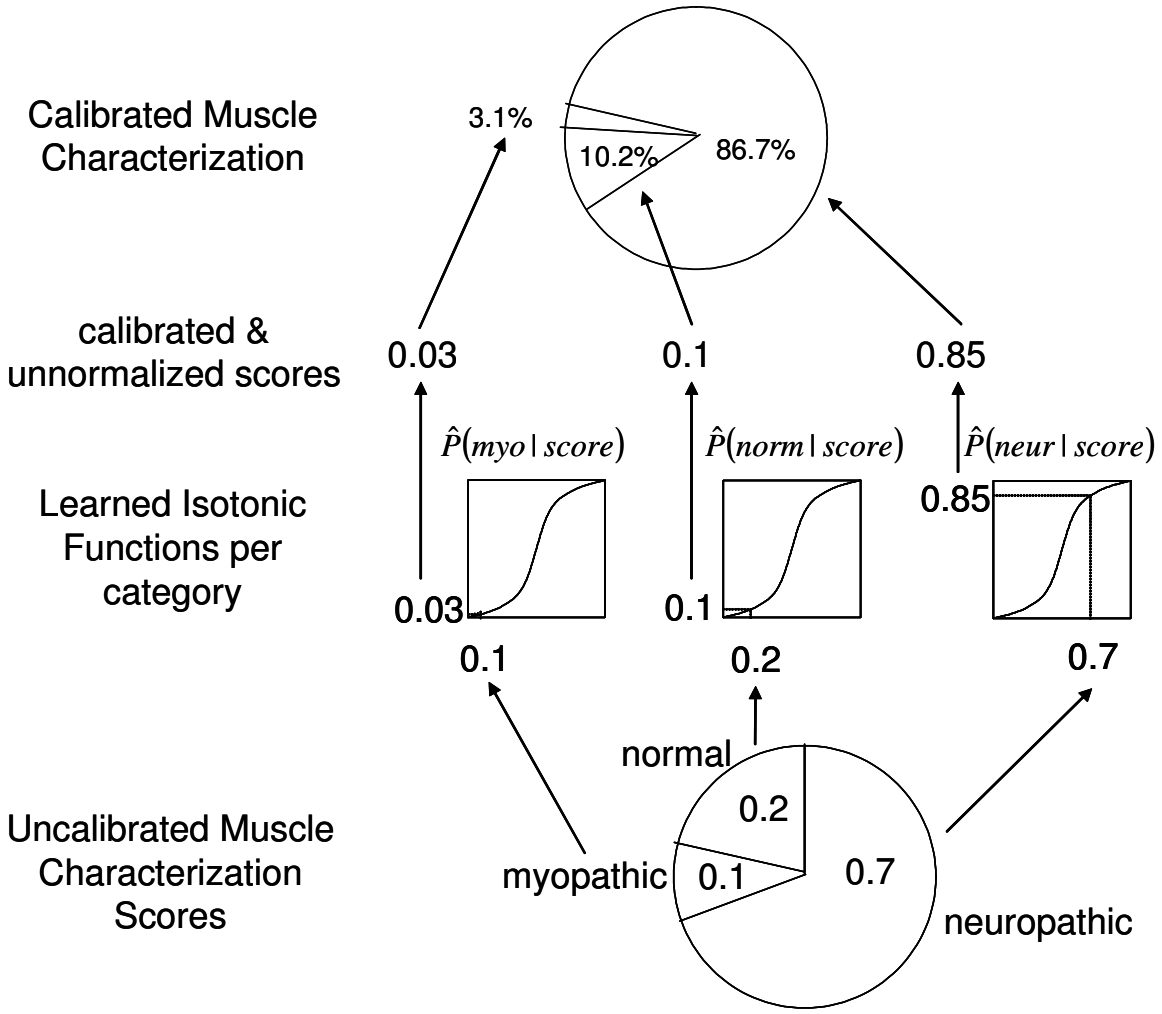


Figure 6.1 Example of Calibrating Muscle Characterization Scores

6.3.1 Generating Reliability Curves

Three hundred different runs of training data were created by randomly choosing 450 out of 500 simulated MUPs per run per category to produce 135,000 MUP characterizations per category. The MUP data was characterized using PD to produce three conditional probabilities for each MUP. In total, 405,000 MUP characterizations were generated. Equal number of virtual muscle data per category was created by randomly selecting MUP characterizations out of the 135,000 MUP

characterizations. Four different sets of 60,000 virtual muscles per category were created where each set was composed of 1, 2, 5 and 10 MUPs respectively per virtual muscle. The sets that had 1 and 2 MUPs per virtual muscle were generated without replacement from the pool of MUP characterizations as 135,000 MUP characterizations is greater than 2 x 60,000. However, sets that had 5 and 10 MUPs per virtual muscle were sampled with replacement from the pool of MUP characterizations. Reliability diagrams of average muscle characterization were compared to Z-Transform muscle characterizations for each of the four sets and per category.

The same process described in the previous paragraph was repeated for the remaining MUP data sets described in Chapter 3 and was done by randomly choosing approximately 90% of the available MUPs per training run and repeating until 135,000 MUP characterizations per category were generated. Four sets of 60,000 virtual muscles per category were generated for each MUP data set where each set was composed of 1, 2, 5 and 10 MUPs respectively per virtual muscle. Reliability diagrams for virtual muscles with greater than 10 MUPs were not drawn because of the tendency of the reliability curves to saturate into sharp sigmoid shapes resembling the shape just before, during, and after the rising edge of a square wave.

Reliability diagrams were drawn by using equation 6.2 for the un-calibrated and calibrated scores. The mean square error of un-calibrated scores denoted MSE_{uncal} and calibrated scores denoted MSE_{cal} were calculated across all N points in a reliability diagram by squaring the difference of each category membership probability to its score and dividing by the number of points as follows:

$$MSE_{uncal}(y_k) = \frac{\sum_{i=1}^{N} \left(P(y_k \mid s_{ki}) - s_{ki}\right)^2}{N} \qquad (6.3)$$

$$MSE_{cal}(y_k) = \frac{\sum_{i=1}^{N} \left(P(y_k \mid \hat{s}_{ki}) - \hat{s}_{ki}\right)^2}{N} \qquad (6.4)$$

6.4 Calibrated Muscle Characterization Results

Un-calibrated MSE increases as the number of MUPs per virtual muscle increases regardless of the data set studied as shown in Tables 6.1 to 6.5. A line fitted through the un-calibrated data points in the reliability diagrams resembles a sigmoid shape where the slope of the middle section of the sigmoid increases as the number of MUPs per virtual muscle increases. Regardless of the category the MSE for every MUP data set becomes smaller for calibrated data in the reliability diagrams as compared to the un-calibrated data.

The MSE of the ZT muscle characterization method did not differ a great deal from the MSE of the AMC method for MUP data sets that had two categories (LHS, EAS and LJUB-Biceps) as shown in Tables 6.1 to 6.3. A line fitted through the points in the reliability diagrams shows that both the calibrated AMC and ZT methods approximately follow the y=x line. Figures 6.2 and 6.3 show the reliability diagrams for neuropathic data using the AMC method for the EAS and LHS MUP data respectively.

The MSE of the un-calibrated and calibrated ZT muscle characterization method was usually less than the MSE of the un-calibrated and calibrated AMC method for MUP data sets that had three categories (simulated and Rigs) for virtual muscles with 5 and 10 MUPs as shown in Tables 6.4 and 6.5. For the simulated MUP data the MSE of the calibrated ZT method averaged about 60% and 42% of the MSE of the calibrated AMC method across the categories for 5 and 10 MUPs per virtual muscle respectively for the simulated MUP data. A line fitted through the points in the reliability diagrams showed that the ZT method follows closer to the y=x line than the AMC method when comparing results for the same number of MUPs per virtual muscle for the simulated data. For the Rigs MUP data the MSE of the calibrated ZT method was less than the calibrated AMC method

except for the Neuropathic category. Figures 6.4 and 6.5 show the reliability diagrams for the calibrated AMC and ZT methods for the Rigs myopathic and neuropathic categories respectively. As shown in Table 6.4 the greatest decrease in MSE of the ZT method as compared to the AMC method for the calibrated neuropathic scores occurred when using 10 MUPs per virtual muscle. The majority of the points as shown in Figure 6.5 for virtual muscles of 5 and 10 MUPs are clustered towards 0 or 1 for both the calibrated AMC and ZT methods thus not leaving a great deal of points in between to get a good sampling of the range of scores.

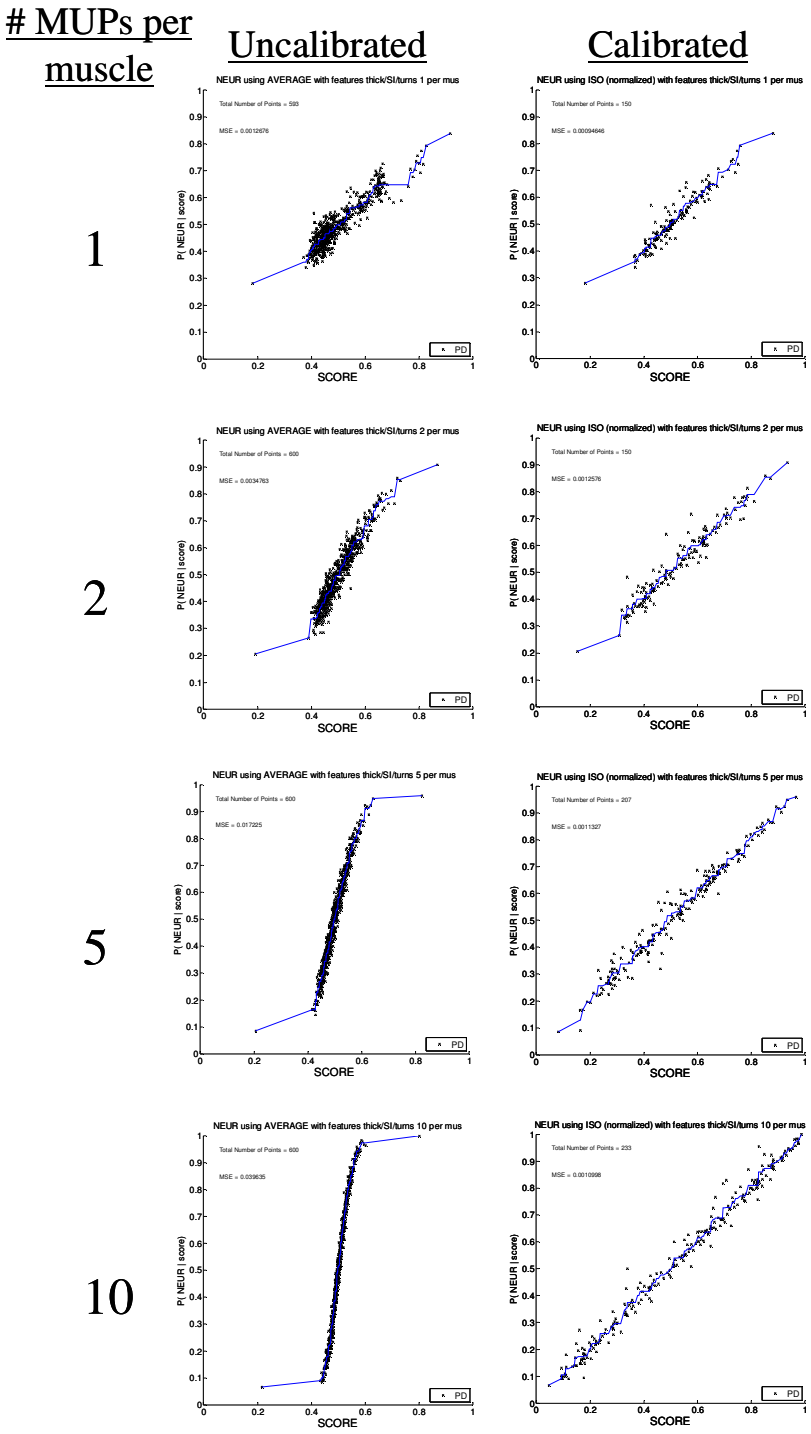


Figure 6.2 Reliability Diagrams for Un-calibrated & Calibrated Neuropathic EAS MUP data Characterizations using AMC

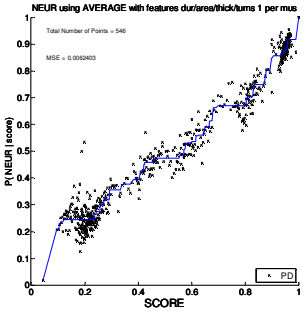
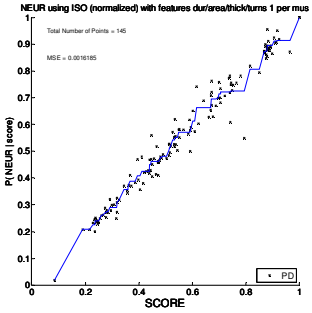
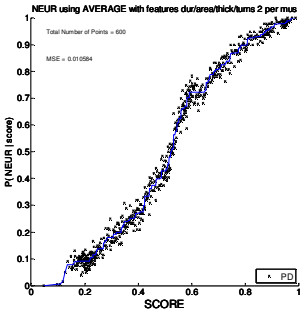
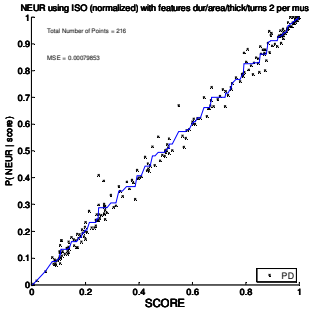
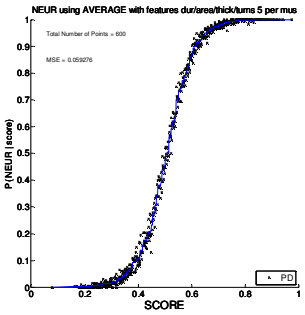
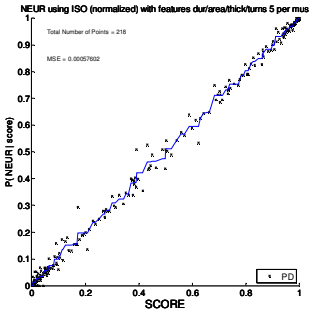
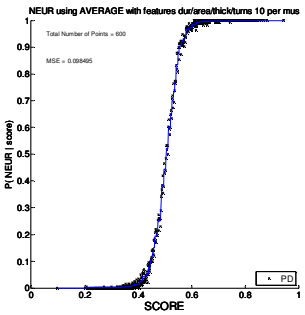
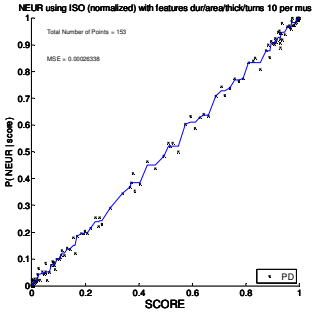
# MUPs per muscle	Uncalibrated	Calibrated
1		
2		
5		
10		

Figure 6.3 Reliability Diagrams for Un-calibrated & Calibrated Neuropathic LHS MUP Data Characterizations using AMC

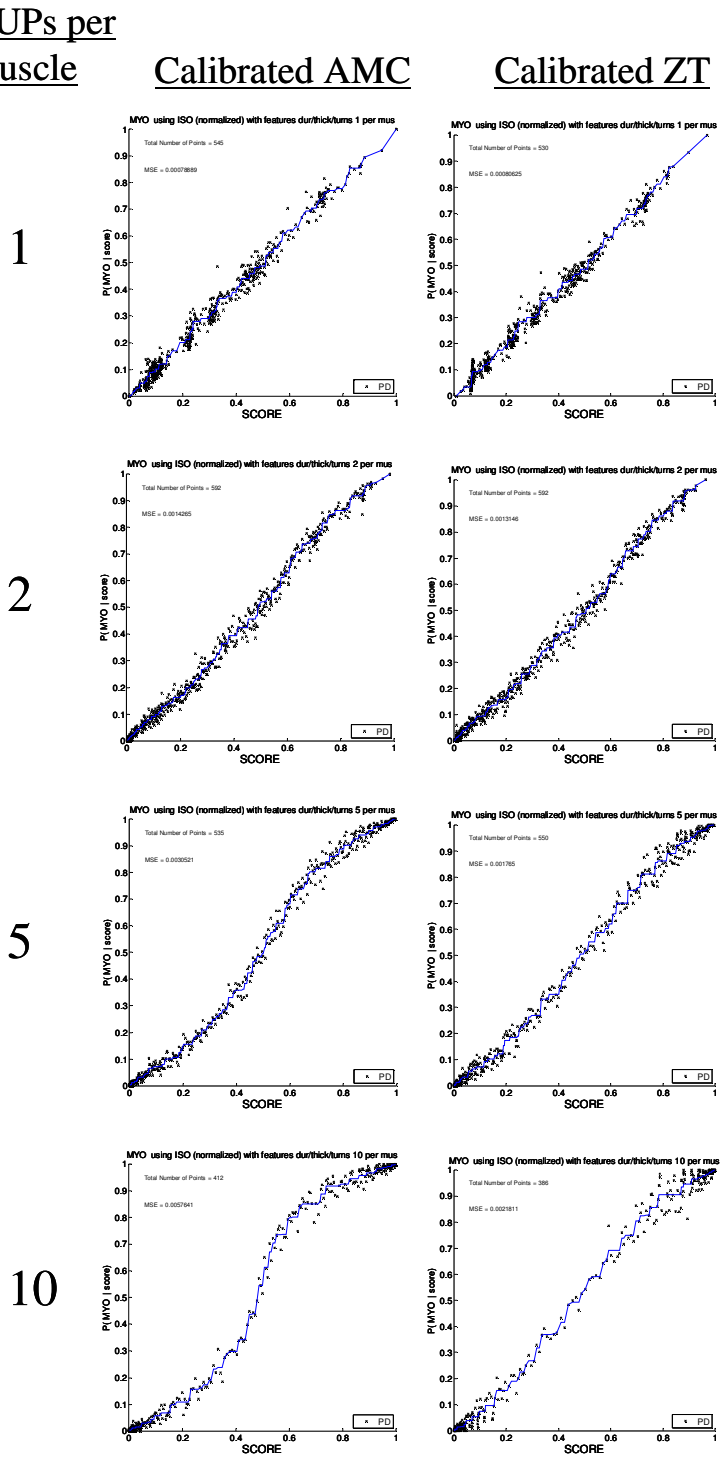


Figure 6.4 Reliability Diagrams for Calibrated Myopathic Rigs MUP Data Characterizations using AMC & ZT

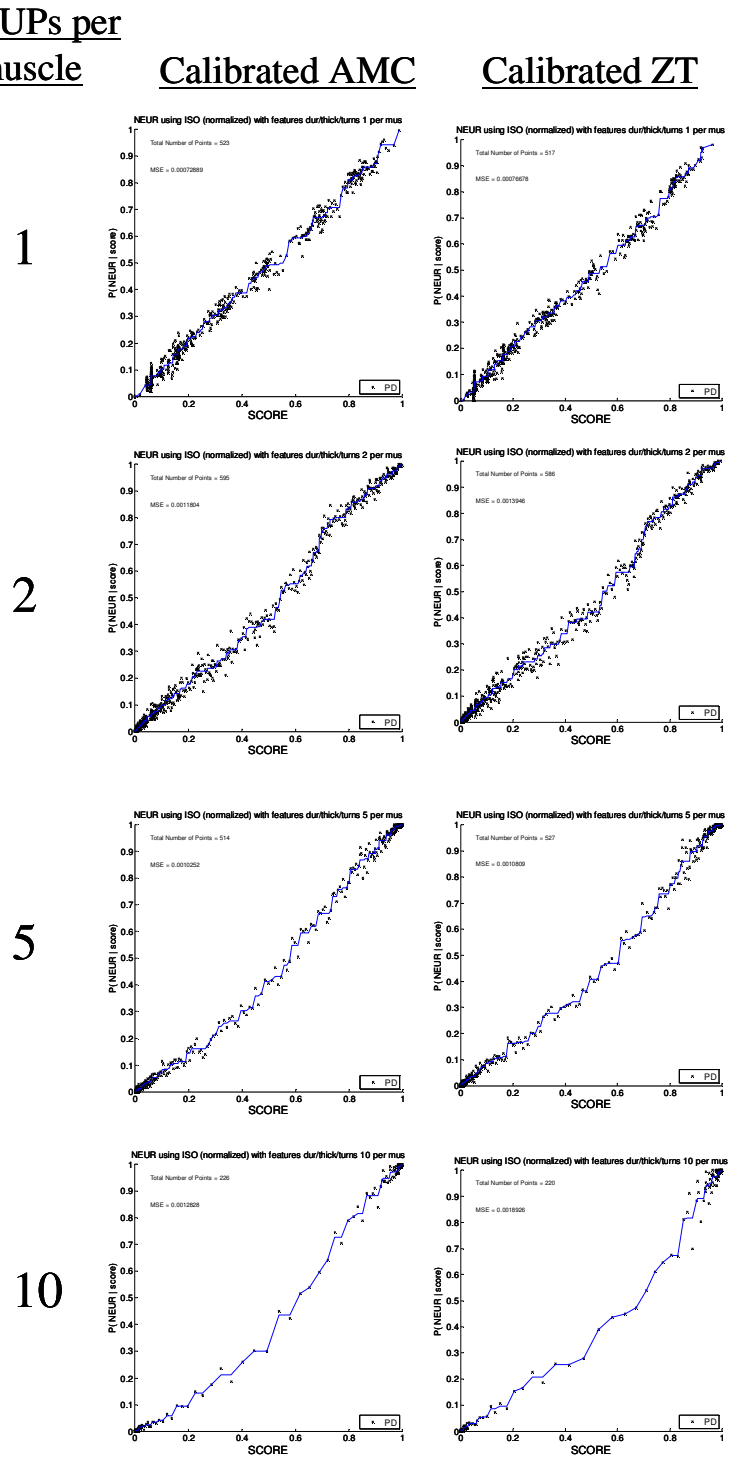


Figure 6.5 Reliability Diagrams for Calibrated Neuropathic Rigs MUP Data Characterizations using AMC & ZT

Table 6.1 MSE of Calibrated & Un-calibrated LHS MUP Data

	MSE of LHS Data			
	Uncalibrated		Calibrated	
x10E-3	AMC	ZT	AMC	ZT
NORM1	6.9	4.2	1.7	2.2
NORM2	10.6	11.6	0.7	0.7
NORM5	59.3	54.7	0.6	0.6
NORM10	98.5	96.7	0.3	0.3
NEUR1	6.2	3.0	1.6	1.6
NEUR2	10.6	11.6	0.8	0.8
NEUR5	59.3	54.7	0.6	0.6
NEUR10	98.5	96.7	0.3	0.3

Table 6.2 MSE of Calibrated & Un-calibrated EAS MUP Data

	MSE of EAS Data			
	Uncalibrated		Calibrated	
x10E-3	AMC	ZT	AMC	ZT
NORM1	1.2	1.1	0.8	0.8
NORM2	3.5	4.0	1.6	1.6
NORM5	17.2	19.7	1.0	1.0
NORM10	39.6	46.5	1.0	1.0
NEUR1	1.3	1.2	0.9	1.0
NEUR2	3.5	4.0	1.3	1.3
NEUR5	17.2	19.7	1.1	1.1
NEUR10	39.6	46.5	1.1	1.1

Table 6.3 MSE of Calibrated & Un-calibrated Ljubljana Biceps MUP Data

	MSE of LJUB-Biceps Data			
	Uncalibrated		Calibrated	
x10E-3	AMC	ZT	AMC	ZT
MYO1	3.8	2.8	1.9	1.9
MYO2	11.3	12.4	0.7	0.7
MYO5	61.6	61.3	0.6	0.6
MYO10	108.5	118.6	0.3	0.3
NORM1	3.7	2.7	1.1	1.1
NORM2	11.3	12.3	1.2	1.2
NORM5	61.6	61.3	0.5	0.5
NORM10	108.5	118.6	0.6	0.6

Table 6.4 MSE of Calibrated and Un-calibrated Simulated MUP Data

	MSE SIMULATED DATA			
	Uncalibrated		Calibrated	
x 10E-3	AMC	ZT	AMC	ZT
MYO1	2.2	2.5	1.2	1.4
MYO2	15.9	14.2	1.5	1.3
MYO5	57.2	29.2	0.9	0.6
MYO10	78.8	28.2	1.0	0.4
NORM1	1.5	2.9	1.1	1.2
NORM2	16.6	14.0	1.4	1.5
NORM5	66.8	37.8	1.7	0.9
NORM10	103.1	44.6	2.1	0.8
NEUR1	3.0	3.5	1.1	1.3
NEUR2	13.6	12.4	1.2	1.1
NEUR5	52.8	25.0	1.2	0.7
NEUR10	76.7	24.6	2.1	1.0

Table 6.5 MSE of Un-calibrated & Calibrated Rigs MUP Data

	MSE Rigs Data			
	Uncalibrated		Calibrated	
x 10E-3	AMC	ZT	AMC	ZT
MYO1	1.0	2.2	0.8	0.8
MYO2	10.4	9.1	1.4	1.3
MYO5	47.3	26.3	3.1	1.8
MYO10	80.3	43.0	5.8	2.2
NORM1	1.2	0.9	0.7	0.5
NORM2	10.1	7.7	1.2	1.3
NORM5	51.4	32.0	2.7	1.7
NORM10	92.9	55.2	4.5	1.8
NEUR1	2.6	2.2	0.7	0.8
NEUR2	11.5	9.7	1.2	1.4
NEUR5	51.6	21.4	1.0	1.1
NEUR10	73.5	20.8	1.3	1.9

6.5 Discussion

Based on the results the AMC or ZT method for muscle characterization are recommended. The ZT method achieves the lowest MSE scores for the three category data sets and equals the performance

of the AMC method for two category data sets. The neuropathic category of the Rigs MUP data was the only data where the ZT method performed slightly worse than the AMC method. The author of this thesis believes that this is due to the polarized clustering of points by the ZT to be near zero or 1 so that not enough points in between 0 and 1 remained to get a robust measure of MSE.

The method presented in this chapter appears to calculate calibrated scores that are reliable, i.e. reflective of their true underlying conditional probabilities. A clinician would have greater trust in a system that produces reliable conditional probabilities fulfilling an important requirement of a CDSS – Section 2.3 under "Report Confidence". A clinic could establish a threshold of conditional probability that needs to be exceeded before declaring an abnormality. Calibrated conditional probability could be quantized into "possible", "probable", or "definite" levels of abnormality useful information for helping a clinician decide whether further testing is required or whether treatment can be applied. In addition, calibrated conditional probabilities could be an indication of the level of involvement of a disorder and could be used to examine the effectiveness of treatments in longitudinal studies. A system that produces reliable conditional probabilities is also transparent in that the system reports for each muscle characterization its level of confidence.

The low MSE scores shown in Tables 6.1 to 6.5 for un-calibrated scores of muscle characterizations using 1 MUP and the upper left reliability diagrams of Figure 6.2 and 6.3 show that the un-calibrated PD scores are well calibrated by how close the reliability diagram follows the y=x line. The un-calibrated scores of muscle characterizations using greater than 1 MUP are not reflective of their underlying conditional probabilities as shown by the increasingly sharper sigmoid shapes and higher MSE values as the number of MUPs used for muscle characterization increases.

The un-calibrated reliability diagrams in Figure 6.2 show the effect of data that does not have a great deal of separation between categories. Because of the ambiguity of the EAS MUP data set there were very few scores lower than 0.4 and scores greater than 0.7. For data sets with good separation

between categories there appears to be a more uniform distribution of points across all values of scores as shown in the un-calibrated column of Figure 6.3.

It appears from the results that the performance of calibrating three category data sets (simulated and Rigs) is poorer than two category data sets (EAS, LHS and LJUB-Biceps) especially as the number of MUPs per muscle increases. The method proposed in this chapter considers one category at a time to draw binary reliability diagrams and then learning an isotonic function. It seemed to work well for the two category data sets because the mapping occurs between one-dimensional spaces. In reality, the three category case is a mapping of a two-dimensional space to another two-dimensional space. Breaking the three category problem into separate one-dimensional problems and the normalization has likely led to a loss of information and therefore poorer results than for the two category data sets.

6.5.1 Application

The following is an initial proposal as to how a clinic conducting EMG examinations could provide calibrated muscle characterizations. A clinic would have control over their labeled training data. Examinations using the CDSS would be consistent with their definition of abnormality.

Training Phase

1. Collect MUP data from patients visiting the clinic.
2. Patients with confirmed diagnosis could have their MUP data labeled as myopathic or neuropathic.
3. Confirmed healthy subjects are used to establish normal data to have a definition of 100% normal. It is recommended that subjects referred to the clinic for neuromuscular symptoms and then cleared by the clinic as being healthy are not to be used to establish normal training data as their data may not be representative of a healthy subject.

4. Features extracted from MUPs are pooled into the following sets: myopathic, normal and neuropathic.
5. Random selection as described in Section 6.3.1 with the pooled MUP data (step 4 above) is used to generate a reliability curve for each category and for the following numbers of MUPs per virtual muscle $n \quad \varepsilon \quad \{1,2,\ldots,9,10\}$.
6. An isotonic function is learned from the reliability curves generated in step 5 leading to thirty learned functions (3 categories x 10 different sets of *n MUPs/virtual muscle*).

Clinical Examination of a Patient

1. A muscle from a patient under examination is characterized using the ZT-PD method. The number of MUPs n detected and used for characterization is noted.
2. The set of isotonic functions learned during the training phase that correspond to n MUPs are used to obtain a calibrated muscle characterization.

There are unresolved issues with the process discussed above. Usually greater than 10 MUPs will be detected from a muscle under examination. Drawing reliability curves for greater than 10 MUPs is not recommended because they saturate into a sharp sigmoid shape with a slope of the rising edge approaching infinity. This saturated sigmoid shape would lead to calibrated scores of either 0, indeterminate (for portion approaching infinite slope) or 1. The MSE of reliability diagrams of calibrated scores cannot be accurately determined because the calibrated scores are clustered at either 0 or 1 and the entire range of scores is not sufficiently sampled to obtain a statistically robust MSE (see reliability diagram bottom right of Figure 6.3 for an example).

Here are some possible strategies for doing calibrated muscle characterizations when the number of MUPs detected per muscle is greater than 10. The following strategies need to be tested for

accuracy as in chapter 5 and for the MSE of their calibrated reliability curves as described in this chapter.

- Use only 10 randomly selected MUPs to use for calibration or
- Rank the MUP characterizations by their MUP conditional probabilities of the category with the highest score. Divide the ordered set of MUP characterizations into sets of 10 and calibrate each set. Average the calibrated scores across these sets to produce an overall calibrated muscle characterization or
- Use the functions learned for $n = 10$ to calibrate any raw muscle characterization score where $n > 10$ or
- Generate reliability curves for all values of n up to a upper limit say 60 and use the appropriate learned function.

Chapter 7
Conclusions

This work recommends using PD based MUP characterizations and transforming a set of MUP characterizations into a muscle characterization using the AMC or Z-transform method. Finally, AMC or the ZT transform can be calibrated by learning an isotonic function for each category using the PAV algorithm and then converting the score into a calibrated conditional probability of a neuromuscular category given the MUPs detected from a muscle.

The AMC-PD or ZT-PD method can be directly used to support clinical decisions related to initial diagnosis as well as treatment and management over time. Decisions are based on facts and not impressions giving electromyography a more reliable role in the diagnosis, management, and treatment of neuromuscular disorders. AMC-PD or ZT-PD based muscle characterization can help make electrophysiological examinations more accurate and objective.

7.1 Success in Meeting the Requirements

This work cannot completely answer the question: does the CDSS improve a clinician's performance? This question is more important than determining the accuracy of the characterizations provided by a CDSS as a stand-alone system and is addressed in the Future work section of this chapter. The function of the CDSS is to provide a characterization. It is the role of the clinician to transform the characterization into a categorization. Currently clinicians use mostly manual methods to recognize patterns created by MUP features and MU activation. The acceptance of a CDSS by clinicians depends on its ability to provide a new point of view that reveals rules or relationships easily seen as well as not easily seen by their current methods.

The section will provide conclusions on the recommended CDSS design against the requirements listed in Section 2.3.

7.1.1 Transparency

MUP characterization using PD was found to be transparent because it explained its characterizations by reporting sets of feature values along with a strength-of-evidence measure supporting or refuting its characterizations. Calibrated muscle characterizations methods in this work can explain their characterizations by ranking the MUPs as sorted by conditional probability of the category with the highest characterization measure. Muscle characterizations were not as transparent as individual MUP characterizations because muscle characterizations require a large amount of information to be integrated into an underlying explanation, see for example Figure 5.2.

7.1.2 Accuracy

AMC-PD based muscle characterizations had higher sensitivity and specificity than conventional muscle characterization methods. It uses the same information available during qualitative examination except it produces consistent, objective evidence while not being dependent on intuition, and biased by other sources of information. As a decision support system, it can aid an electromyographer in a single but important step of an EMG examination and can provide an objective record over time that facilitates longitudinal studies.

7.1.3 Confidence

According to the reliability curves in Chapter 6, the AMC-PD or ZT-PD method can be calibrated into a probability estimate that corresponds to confidence. Clinicians can understand the reliability of an individual characterization and perhaps use confidence thresholds to declare that a patient has an abnormality, e.g. only calibrated probability estimates of greater than 90% should confirm the

presence of a category of abnormality and estimates of less than 90% require further testing or revision of the hypothesis as to the underlying cause of a patient's symptoms.

7.1.4 Generalization

The methods for training and testing were done using jackknifing across a number of different trials. As well, several different MUP data sets were studied. The AMC-PD and hence the ZT-PD method is accurate across different MUP data sets - even for data such as the EAS MUP data set that does not have a great deal of separation between the categories.

7.1.5 Handle Missing Data

The effect of missing data was not studied in this work.

7.2 Research Contributions

In the author's opinion, the original contributions of this work are:

1. Methods for doing PD based MUP characterizations were implemented and evaluated.
2. Methods for AMC-PD based muscle characterizations were developed and evaluated. The AMC-PD method exceeds the performance of conventional muscle characterization techniques as per the requirements of a CDSS.
3. The PAV method for calibrating the AMC-PD and ZT-PD muscle characterization methods were implemented and evaluated.
4. An equation for transforming WOE into conditional probabilities was derived.
5. The thesis suggests that different muscle types can be pooled together for training and testing data and found to provide reasonable accuracy.
6. A graphical method of displaying MUP characterizations in a transparent manner was developed.

7.3 Future Work

7.3.1 Graphical User Interface (GUI)

The methods developed and recommended in this work should be built into a system capable of being used in a clinical setting. Feedback about the usability and effectiveness of CDSS prototypes with the developed GUI should be sought from clinicians that conduct EMG examinations.

7.3.2 ZT-PD Method

Other MUP data sets that have three or more categories should be examined as to the reliability diagrams produced by the AMC and ZT methods. The ZT method may provide better calibrated conditional probabilities than the AMC method in terms of low MSE score in a reliability diagram as the number of categories to be characterized increases.

If the ZT method does not perform well for greater than three categories then finding a method that can perform well for higher dimensional calibration may be a worthwhile research project.

7.3.3 Transparent Muscle Characterization

Although a demonstration of the transparency of MUP characterizations was provided, a method that can provide a transparent muscle characterization would be useful.

7.3.4 Correlation to Level of Involvement

The characterization measures of the AMC method showed excellent correlation to level of involvement of neuropathic and myopathic neuromuscular disorders when analyzing electromyographic signals based on the simulated MUP data described in this thesis[66]. Verifying that the calibrated conditional probability estimates produced are correlated with the level of

involvement of a disorder for clinical data would be useful. This is challenging because it is difficult to accurately know the level of a disorder for a real muscle.

7.3.5 Addition of Other Electrophysiological Features

Other features should be considered and evaluated. Both new shape and or size based features as well as features derived from firing patterns of MUs. As well, integrating results from other QEMG methods into an overall system may be possible and useful.

7.3.6 Examination of Accuracy versus Number of MUPs

It would be worthwhile to examine the effect of using less than the total number of MUPS detected to determine a muscle characterization and to see the effect of reducing the number of MUPs on accuracy. This may allow clinicians to complete examinations more quickly if they can obtain an accurate muscle characterization with fewer MUPs.

7.3.7 Clinical Evaluation

Friedman et al. describes the results of a trial on how two different CDSS systems enhance or degrade diagnostic reasoning [67]. The study describes a consultation as positive when the clinician made a correct diagnosis after a CDSS consultation but did not include it in their initial diagnosis of the same data. A consultation is negative when the clinician made the correct diagnosis in their initial diagnosis but reversed it after consultation with the CDSS. There should be a net gain of positive consultations versus negative consultations to consider the CDSS as a success in enhancing diagnostic reasoning. A trial like one described by Friedman et al. is recommended as future work.

Appendix A

Derivation of Compound Rule Conditional Probability for PD

MUP = y_k MUP was detected in a muscle of category y_k

MUP $\neq$ y_k MUP was <u>not</u> detected in a muscle of category y_k

x_k^* compound rule associated with category y_k

WOE Weight of Evidence

Definition of WOE.

$$WOE = \log_2 \frac{P(x_k^* \mid MUP = y_k)}{P(x_k^* \mid MUP \neq y_k)} \qquad \text{(A.1)}$$

Definition of conditional probability.

$$P\left(x_k^* \mid MUP = y_k\right) = \frac{P\left(x_k^*, MUP = y_k\right)}{P\left(MUP = y_k\right)} \qquad \text{(A.2)}$$

Total Law of Probability.

$$P\left(x_k^*\right) = P\left(x_k^* \mid MUP = y_k\right) \cdot P\left(MUP = y_k\right) + P\left(x_k^* \mid MUP \neq y_k\right) \cdot P\left(MUP \neq y_k\right) \qquad \text{(A.3)}$$

Rearrange A.3

$$P\left(x_k^* \mid MUP \neq y_k\right) = \frac{P\left(x_k^*\right) - P\left(x_k^* \mid MUP = y_k\right) \cdot P\left(MUP = y_k\right)}{P\left(MUP \neq y_k\right)} \qquad \text{(A.4)}$$

Sum of all prior probabilities equals 1.

$$P\left(MUP = y_k\right) + P\left(MUP \neq y_k\right) = 1 \qquad \text{(A.5)}$$

Sub A.5 into A.4

$$P\left(x_k^* \mid MUP \neq y_k\right) = \frac{P\left(x_k^*\right) - P\left(x_k^* \mid MUP = y_k\right) \cdot P\left(MUP = y_k\right)}{1 - P\left(MUP = y_k\right)} \qquad \text{(A.6)}$$

Sub A.2 into A.6

$$P\left(x_k^* \mid MUP \neq y_k\right) = \frac{P\left(x_k^*\right) - P\left(x_k^*, MUP = y_k\right)}{1 - P\left(MUP = y_k\right)} \qquad \text{(A.7)}$$

Sub A.2 and A.7 into A.1

$$WOE = \log_2 \frac{P\left(x_k^*, MUP = y_k\right) \cdot \left(1 - P\left(MUP = y_k\right)\right)}{P\left(MUP = y_k\right) \cdot \left(P\left(x_k^*\right) - P\left(x_k^*, MUP = y_k\right)\right)} \qquad \text{(A.8)}$$

$$Now \quad letting \quad \Phi = \frac{1 - P(MUP = y_k)}{P(MUP = y_k)} \qquad \text{(A.9)}$$

Sub A.9 into A.8

$$WOE = \log_2 \left(\frac{P\left(x_k, MUP = y_k\right) \cdot \Phi}{P\left(x_k^*\right) - P\left(x_k, MUP = y_k\right)} \right) \qquad \text{(A.10)}$$

Divide numerator and denominator of A.10 by $P\left(x_k, MUP = y_k\right)$

$$WOE = \frac{\Phi}{\frac{P\left(x_k^*\right)}{P\left(x_k^*, MUP = y_k\right)} - 1} \qquad \text{(A.11)}$$

but $P\left(MUP = y_k \mid x_k^*\right) = \frac{P\left(x_l^*, MUP = y_k\right)}{P\left(x_k^*\right)}$, so substituting into A.11

$$WOE = \log_2 \left(\frac{\Phi}{\frac{1}{P\left(MUP = y_k \mid x_l^*\right)} - 1} \right) \qquad \text{(A.12)}$$

rearranging and substituting Φ *back in, so* $\therefore$

$$P\left(MUP = y_k \mid x_k^*\right) = \frac{1}{\left[\left(2^{-WOE}\right) \cdot \left(\frac{1 - P\left(MUP = y_k\right)}{P\left(MUP = y_k\right)}\right)\right] + 1} \qquad \text{(A.13)}$$

Appendix B
Derivation of Bayes Rule for Multiple Pieces of Evidence

In both (Pfeiffer. 1999) and this paper, we are using Bayes theorem for combining the evidence of multiple MUPs. All of the feature values of a single MUP can be thought of as a single piece of evidence *e*. Pfeiffer used the recursive form of Bayes theorem which is used to update the posterior probability as new evidence appears. In this paper, the form of Bayes formulae that is used is called "Bayes Theorem for Multiple Pieces of Evidence" and is identical to the recursive form of Bayes theorem when all pieces of evidence have been collected and the recursion is finished. Below shows how Bayes theorem is derived for multiple pieces of evidence under the assumption that each piece of evidence is conditionally independent of the other pieces of evidence. First, we start with Bayes theorem for a single piece of evidence and multiple categories. More information about Bayes Theorem for multiple pieces of evidence can be found in:

Duda R, Hart PE. Pattern Classification. 2nd edition. John Wiley and Sons, Inc. 2001.

Bayes Theorem – One piece of evidence & Multiple Categories

$Y_i = \text{Category } i,\ i \in \{1,\ldots,K\}$

e = Single piece of evidence

$$P(Y_i \mid e) = \frac{P(e \mid Y_i) \cdot P(Y_i)}{P(e)} \qquad \text{(B.1)}$$

Total Law of Probability assuming each category forms a disjoint set, e.g. a muscle in the training data can only belong to one category.

$$P(e) = \sum_{i=1}^{K} P(e \mid Y_i) P(Y_i) \qquad \text{(B.2)}$$

$$\therefore \quad P(Y_i \mid e) = \frac{P(e \mid Y_i)\,P(Y_i)}{\sum_{i=1}^{K} P(e \mid Y_i) P(Y_i)} \qquad \text{(B.3)}$$

Bayes Theorem – Multiple pieces of evidence & Multiple Categories

Y_i = Category i, $i \in \{1,\ldots,K\}$

e_j = Piece of evidence j, $j \in \{1,\ldots,n\}$

$$P(Y_i \mid e_1,\ldots,e_n) = \frac{P(e_1,\ldots,e_n \mid Y_i)\cdot P(Y_i)}{P(e_1,\ldots,e_n)} \qquad \text{(B.4)}$$

Total Law of Probability assuming each category forms a disjoint set.

$$P(e_1,\ldots,e_n) = \sum_{i=1}^{K} P(e_1,\ldots,e_n \mid Y_i) P(Y_i) \qquad \text{(B.5)}$$

$$\therefore \quad P(Y_i \mid e_1,\ldots,e_n) = \frac{P(e_1,\ldots,e_n \mid Y_i)\,P(Y_i)}{\sum_{i=1}^{K} P(e_1,\ldots,e_n \mid Y_i) P(Y_i)} \qquad \text{(B.6)}$$

If we assume conditional independence of the evidence, i.e. the pieces of evidence are statistically independent given a category then:

$$P(e_1, e_2,\ldots,e_n \mid H_i) = P(e_1 \mid H_i)\,P(e_2 \mid H_i)\cdot\ldots\cdot P(e_n \mid H_i) \qquad \text{(B.7)}$$

so

$$P(Y_i \mid e_1, e_2,\ldots,e_n) = \frac{\left[P(e_1 \mid Y_i)\,P(e_2 \mid Y_i)\cdot\ldots\cdot P(e_n \mid Y_i)\right]P(Y_i)}{\sum_{i=1}^{K}\left\{\left[P(e_1 \mid Y_i)\,P(e_2 \mid Y_i)\cdot\ldots\cdot P(e_n \mid Y_i)\right]P(Y_i)\right\}} \qquad \text{(B.8)}$$

Equation B.8 above shows the form of Bayes theorem for multiple pieces of evidence when the pieces of evidence are statistically independent given a category.

Derivation of Bayes Rule for Muscle Characterization

Start with equation B.8 and substitute in muscle category and MUPs for pieces of evidence.

$$P(MUS = y_k \mid MUP_1, MUP_2, \ldots MUP_n) = \frac{P_0(y_k)\prod_{i=1}^{N} P(MUP_i \mid y_k)}{\sum_{j=1}^{K}\left[P_0(y_j)\prod_{i=1}^{N} P(MUP_i \mid y_j)\right]} \quad \text{(B.9)}$$

Rearrange Bayes rules for a single piece of evidence

$$P(MUP_i \mid y_k) = \frac{P(y_k \mid MUP_i)\cdot P(MUP_i)}{P_0(y_k)} \quad \text{(B.10)}$$

Sub B.10 into B.9

$$P(MUS = y_k \mid MUP_1, MUP_2, \ldots MUP_n) = \frac{P_0(y_k)\prod_{i=1}^{N} \frac{P(y_k \mid MUP_i)\cdot P(MUP_i)}{P_0(y_k)}}{\sum_{j=1}^{K}\left[P_0(y_j)\prod_{i=1}^{N} \frac{P(y_j \mid MUP_i)\cdot P(MUP_i)}{P_0(y_j)}\right]} \quad \text{(B.11)}$$

Assume that all prior probabilities are equal, i.e. $P_0(y_1) = P_0(y_2) = \ldots = P_0(y_K) = 1/K$

Then B.11 reduces to B.12. Each MUP is assumed to provide independent information

$$P(MUS = y_k \mid MUP_1, MUP_2, \ldots MUP_n) = \frac{\prod_{i=1}^{N} P(y_k \mid MUP_i)}{\sum_{j=1}^{K}\left[\prod_{i=1}^{N} P(y_j \mid MUP_i)\right]} \quad \text{(B.12)}$$

Appendix C
Saturation of Bayes' Muscle Characterization

Appendix C shows that the Bayesian muscle characterization technique converges to 1 or 0 as the number of MUPs used for characterization increases.

Two category $\{y_1, y_2\}$ muscle categorization will be considered.

Say that we start with the following MUP characterization for a muscle:

$$P(y_1 \mid MUP_1) = 0.5 + \varepsilon \qquad \text{(C.1)}$$

$$P(y_2 \mid MUP_1) = 0.5 - \varepsilon \qquad \text{(C.2)}$$

N MUPs are found such that they all have the same MUP characterizations:

$$P(y_1 \mid MUP_1) = P(y_1 \mid MUP_2) = \ldots = P(y_1 \mid MUP_N) = 0.5 + \varepsilon \quad \text{(C.3)}$$

$$P(y_2 \mid MUP_1) = P(y_2 \mid MUP_2) = \ldots = P(y_2 \mid MUP_N) = 0.5 - \varepsilon \quad \text{(C.4)}$$

Using Equation 5.4 and setting its denominator = den

$$P(y_1 \mid \{MUPs\}) = \frac{P(y_1 \mid MUP_1)^N}{den} = \frac{(0.5+\varepsilon)^N}{den} \qquad \text{(C.5)}$$

$$P(y_2 \mid \{MUPs\}) = \frac{P(y_2 \mid MUP_1)^N}{den} = \frac{(0.5-\varepsilon)^N}{den} \qquad \text{(C.6)}$$

Consider the ratio of C.5 to C.6 to consider how quickly one conditional probability grows with respect to the other as N increases.

$$R = \left(\frac{0.5 + \varepsilon}{0.5 - \varepsilon} \right)^N \qquad \text{(C.7)}$$

Equation C.7 shows that R approaches ∞ as N approaches ∞ for positives values of ε. This means that C.5 approaches 1 and C.6 approaches 0.

Appendix D

Proof that Z-transform Provides Same Ranking as Averaging

Appendix D is a proof that shows that the Z transform of MUP characterizations has the same ranking of categories as sorted by their muscle characterization scores as compared to the AMC-PD method on a per muscle basis.

First the formulae for AMC muscle characterization for category k from equation 5.3.

$$AMC_k = \frac{\sum_{i=1}^{N} P_{ki}}{N} \qquad \text{(D.1)}$$

The Z transform of category k across N MUPs from equation 6.3 and 6.4.

$$Z_k = \frac{\sum_{i=1}^{N} \left(\frac{P_{ki} - \mu}{\sigma} \right)}{\sqrt{N}} \qquad \text{(D.2)}$$

Re-arranging D.2

$$Z_k = \frac{\sum_{i=1}^{N} P_{ki} - \sum_{i=1}^{N} \mu}{\sigma\sqrt{N}} = \frac{\sum_{i=1}^{N} P_{ki} - N\mu}{\sigma\sqrt{N}} \qquad \text{(D.3)}$$

D.3 can be re-written where a, b are constants.

$$Z_k = a\sum_{i=1}^{N} P_{ki} - b \qquad \text{(D.4)}$$

So if: $AMC_1 < AMC_2 < \ldots < AMC_K$

then $Z_1 < Z_2 < \ldots < Z_K$ because D.4 is a linear scaling and shifting of AMC_k. Hence AMC has the same ranking as Z transform across categories for a given muscle characterization.

References

[1] G. Pfeiffer, "The Diagnostic Power of Motor Unit Potential Analysis: An Objective Bayesian Approach," *Muscle & Nerve,* vol. 22, pp. 584-591, 1999.

[2] A. Tversky and D. Kahnemann, "Judgement under uncertainty," *Science,* vol. 185, pp. 1124-1131, 1974.

[3] C. D. Wickens and J. G. Hollands, *Engineering Psychology and Human Performance.* ,Third ed.New Jersey 07458: Prentice-Hall Inc., 2000,

[4] R. Kendall and R. A. Werner, "Interrater reliability of the needle examination in lumbosacral radiculopathy," *Muscle Nerve,* vol. 34, pp. 238-241, August 2006. 2006.

[5] T. L. Stedman, *Stedman's Medical Dictionary.* ,27th edition ed.Philadelphia: Lippincott Williams & Wilkins, 2000, pp. 2098.

[6] S. D. Nandedkar, E. V. Stalberg and D. Sanders, "Quantitative EMG," in *Electrodiagnostic Medicine* ,2nd ed.D. Dumitru, A. A. Amato and M. Zwarts, Eds. Philadelphia, PA: Hanley & Belfus Inc., 2002,

[7] D. W. Stashuk, "Decomposition and quantitative analysis of clinical electromyographic signals," *Medical Engineering and Physics,* vol. 21, pp. 389-404, July, 1999. 1999.

[8] D. W. Stashuk and W. F. Brown, "Quantitative electromyography," in *Neuromuscular Function and Disease* ,1st ed., vol. 1, W. F. Brown, C. F. Bolton and M. J. Aminoff, Eds. Philadelphia, PA: Elsevier Science, 2002, pp. 311-348.

[9] E. Stålberg and J. V. Trontelj, "The study of normal and abnormal neuromuscular transmission with single fibre electromyography," *Journal of Neuroscience Methods,* vol. 74, pp. 145-154, 6/27. 1997.

[10] C. S. Pattichis, C. N. Schizas and L. T. Middleton, "Neural network models in EMG diagnosis," *Biomedical Engineering, IEEE Transactions on,* vol. 42, pp. 486-496, 1995.

[11] G. Pfeiffer and K. Kunze, "Discriminant classification of motor unit potentials (MUPs) successfully separates neurogenic and myogenic conditions. A comparison of multi- and univariate diagnostic algorithms for MUP analysis." *Electroencephalography Clinical Neurophysiology,* vol. 97, pp. 191-207, 1995.

[12] E. Stålberg, C. Bischoff and B. Falck, "Outliers, A Way to Detect Abnormality in Quantiative EMG," *Muscle and Nerve,* vol. 17, pp. 392-399, 1994.

[13] C. R. Stewart, S. D. Nandedkar, J. M. Massey, J. M. Gilchrist, P. E. Barkhaus and D. B. Sanders, "Evaluation of an automatic method of measuring features of motor unit action potentials." *Muscle and Nerve,* vol. 12(2), pp. 141-148, 1989 Feb.

[14] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif. Intell. Med.,* vol. 23, pp. 89-109, 08/. 2001.

[15] M. Sprogar, M. Lenic and S. Alayon, "Evolution in Medical Decision Making," *Journal of Medical Systems,* vol. 26, pp. 479-489, October 2002. 2002.

[16] T. M. Strat and J. D. Lowrance, "Explaining evidential analyses," *International Journal of Approximate Reasoning,* vol. 3, pp. 299-353, July, 1989. 1989.

[17] C. Feng and D. Michie, "Machine learning of rules and trees," in *Machine Learning, Neural and Statistical Classification* D. Michie, D. J. Spiegelhalter and C. C. Taylor, Eds. Herfordshire, England: Ellis Horwood Limited, 1994,

[18] K. M. Chan, "Needle EMG abnormalities in neurogenic and muscle diseases," in *Neuromuscular Function and Disease* , vol. 1, W. F. Brown, C. F. Bolton and M. J. Aminoff, Eds. Philadelphia, PA: Elsevier Science, 2002, pp. 359-368.

[19] R. Duda and P. E. Hart, "Pattern Classification. 2nd edition." *John Wiley and Sons, Inc.,* 2001.

[20] J. M. Chavin and W. F. Brown, "Negative signs and symptoms of peripheral nerve and muscle disease," in ,1st ed., vol. 1, W. F. Brown, C. F. Bolton and M. J. Aminoff, Eds. Philadelphia, PA: Elsevier Science, 2002, pp. 369-385.

[21] S. Vingtoft, A. Fugslang-Frederissen and Ronager, J. et al., "KANDID – An EMG Decision Support System – Evaluated in a European Multicenter Trial," *Muscle and Nerve,* vol. 16, pp. 520-529, 1993.

[22] S. J. Russell, P. Norvig and J. F. Canny, *Artificial Intelligence : A Modern Approach.* ,2nd ed.Upper Saddle River, N.J.: Prentice Hall, 2003,

[23] S. Andreassen, B. Falck and K. G. Olesen, "Diagnostic Function of the Microhuman Prototype of the Expert System MUNIN." *Electroencephalography and Clinical Neurophysiology,* vol. 85, pp. 143-157, 1992.

[24] I. H. Witten and F. Eibe, *"Data Mining: Practical Machine Learning Tools and Techniques".* ,2nd ed.San Francisco: Morgan Kaufman, 2005,

[25] D. Szafron, R. Greiner, P. Lu and et al, "TCXplain: Transparent explanation of naïve bayes classifications." University of Alberta, Tech. Rep. Report TR03-09. University of Alberta., 2003.

[26] C. Bull, M. Chiogna, R. Franklin and D. Spiegelhalter, "Expert derived and automatically generated classification trees: An example from pediatric cardiology," in 1993, pp. 217-220.

[27] J. R. Quinlan, "Induction of decision trees." *Machine Learning,* vol. 1, 1986.

[28] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *In Proceedings of the Eighteenth International Conference on Machine Learning,* 2001, pp. 609-616.

[29] M. M. Tatsuoka, "Multivariate Analysis: Techniques for Educational and Psychological Research. 2nd Ed." *New York, NY; Macmillan Publishing Company,* 1998.

[30] A. K. C. Wong and Y. Wang, "Discovery of high order patterns," in *Proceedings of the 1995 IEEE International Conference on Systems, Man and Cybernetics. Part 2 (of 5), Oct 22-25 1995,* 1995, pp. 1142-1147.

[31] A. K. C. Wong and Y. Wang, "High-order pattern discovery from discrete-valued data," *IEEE Trans. Knowled. Data Eng.,* vol. 9, pp. 877-893, 1997.

[32] A. K. C. Wong and Y. Wang, "Pattern discovery: A data driven approach to decision support," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews,* vol. 33, pp. 114-124, 2003.

[33] Y. Wang, "High-Order Pattern Discovery and Analysis of Discrete-Valued Data Sets," *University of Waterloo, PhD Thesis,* 1997.

[34] E. Stalberg, S. D. Nandedkar, D. B. Sanders and B. Falck, "Quantitative Motor Unit Potential Analysis." *Journal of Clinical Neurophysiology,* vol. 13, pp. 401-422, 1996.

[35] S. D. Nandedkar, P. E. Barkhaus, D. B. Sanders and E. V. Stålberg, "Analysis of amplitude and area of concentric needle EMG motor unit action potentials," *Electroencephalogr. Clin. Neurophysiol.,* vol. 69, pp. 561-567, June, 1988. 1988.

[36] M. Sonoo and E. Stalberg, "
The ability of MUP parameters to discriminate between normal and neurogenic MUPs in concentric EMG: analysis of the MUP “thickness” and the proposal of “size index.”," *Electroencephalogr Clin Neurophysiol,* vol. 89, pp. 291-303, 1993.

[37] M. Sonoo, "New attempts to quantify concentric needle electromyography," *Muscle Nerve,* vol. 999, pp. S98-S102, 2002. 2002.

[38] A. Hamilton-Wright and D. W. Stashuk, "Physiologically based simulation of clinical EMG signals," *IEEE Transactions on Biomedical Engineering,* vol. 52, pp. 171-183, 2005.

[39] A. Hamilton-Wright, D. W. Stashuk and T. J. Doherty, "
Simulation of disease effects on muscle structure, activation and acquired electromyography," *Muscle and Nerve,* pp. S126, 2003.

[40] B. R. Brooks, R. G. Miller, M. Swash and T. L. Munsat, "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral
sclerosis," *Amyotroph Lateral Scler Other Motor Neuron Disord,* vol. 1, pp. 293-299, 2000 Dec.

[41] S. G. Boe, D. W. Stashuk and T. J. Doherty, "Motor unit number estimation by decomposition-enhanced spike-triggered averaging: Control data, test-retest reliability, and contractile level effects," *Muscle Nerve,* vol. 29, pp. 693-699, May 2004. 2004.

[42] S. G. Boe, D. W. Stashuk, W. F. Brown and T. J. Doherty, "Decomposition-based quantitative electromyography: Effect of force on motor unit potentials and motor unit number estimates," *Muscle Nerve,* vol. 31, pp. 365-373, March 2005. 2005.

[43] S. G. Boe, D. W. Stashuk and T. J. Doherty, "Within-subject reliability of motor unit number estimates and quantitative motor unit analysis in a distal and proximal upper limb muscle," *Clinical Neurophysiology,* vol. 117, pp. 596-603, March, 2006. 2006.

[44] M. Nikolic, "Detailed Analysis of Clinical Electromyography Signals: EMG Decomposition, Findings and Firing Pattern Analysis in Controls and Patients with Myopathy and Amytrophic Lateral Sclerosis," *University of Copenhagen,* 2001.

[45] E. Stålberg, B. Falck, M. Sonoo, S. Stålberg and M. Åström, "Multi-MUP EMG analysis -- a two year experience in daily clinical work," *Electroencephalography and Clinical Neurophysiology/ Electromyography and Motor Control,* vol. 97, pp. 145-154, June, 1995. 1995.

[46] S. Podnar, D. B. Vodusek and E. Stalberg, "Comparison of quantitative techniques in anal sphincter electromyography," *Muscle Nerve,* vol. 25, pp. 83-92, January 2002. 2002.

[47] S. Podnar, "Criteria for neuropathic abnormality in quantitative anal sphincter electromyography," *Muscle Nerve,* vol. 30, pp. 596, November 2004. 2004.

[48] S. Podnar and D. B. Vodusek, "Standardization of anal sphincter electromyography: Utility of motor unit potential parameters," *Muscle Nerve,* vol. 24, pp. 946-951, July 2001. 2001.

[49] S. Podnar, D. B. Vodusek and E. Stalberg, "Standardization of anal sphincter electromyography: normative data," *Clinical Neurophysiology,* vol. 111, pp. 2200-2207, December, 2000. 2000.

[50] S. Podnar and D. B. Vodusek, "Standardisation of anal sphincter EMG: high and low threshold motor units," *Clinical Neurophysiology,* vol. 110, pp. 1488-1491, August 1, 1999. 1999.

[51] S. Podnar and J. Zidar, "Sensitivity of motor unit potential analysis in facioscapulohumeral muscular dystrophy," *Muscle Nerve,* vol. 34, pp. 451-456, October 2006. 2006.

[52] L. J. Pino, D. W. Stashuk and S. Podnar, "Bayesian characterization of external anal sphincter muscles using quantitative electromyography," *Clinical Neurophysiology,* vol. 119, pp. 2266-2273, 10. 2008.

[53] Y. Wang, "High-Order Pattern Discovery and Analysis of Discrete-Valued Data Sets," 1997.

[54] A. Trujillo-Ortiz, R. Hernandez-Walls and S. Perez-Osuna, "RAFisher2cda:Canonical Discriminant Analysis. A MATLAB file. [WWW document]. " 2004.

[55] E. Stalberg and H. Erdem, "Quantitative motor unit potential analysis in routine." *Electromyogr Clin Neurophysiol.,* vol. 42, pp. 433-442, 2002.

[56] F. Buchthal, *An Introduction to Electromyography.* Copenhagen: Scandinavian University Books, 1957, pp. 43.

[57] M. Swash, "What does the neurologist expect from clinical neurophysiology?" *Muscle Nerve,* vol. 999, pp. S134, 2002. 2002.

[58] A. Hamilton-Wright, D. W. Stashuk and L. J. Pino, "On weight of evidence based reliability in pattern discovery," in 2006,

[59] A. Hamilton-Wright, D. W. Stashuk and L. J. Pino, "Internal measures of reliability in 'pattern discovery' based fuzzy inference." in *Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems,* 2006,

[60] S. Podnar and M. Mrkaic, "Predictive power of motor unit potential parameters in anal sphincter electromyography," *Muscle Nerve,* vol. 26, pp. 389-394, September 2002. 2002.

[61] L. J. Pino, D. W. Stashuk, S. G. Boe and T. J. Doherty, "Motor unit potential characterization using "pattern discovery"," *Medical Engineering and Physics,* vol. 30, pp. 563-73, 06. 2008.

[62] M. Whitlock, "Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach," *J. Evol. Biol.,* vol. 18, pp. 1368-73, Sep. 2005.

[63] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2002, pp. 694-699.

[64] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences.* ,2nd ed ed.Amsterdam ; Boston: Academic Press, 2006, pp. xvii, 627 p.

[65] L. Dumbgen, "Statistical Software for Matlab " *Avai,* 2008.

[66] L. J. Pino and D. W. Stashuk, "Using motor unit potential characterizations to estimate neuromuscular disorder level of involvement," in *30th Annual International IEEE EMBS Conference, Vancouver, British Columbia, Canada, August,* 2008,

[67] C. P. Friedman, A. S. Elstein, F. M. Wolf, G. C. Murphy, T. M. Franz, P. S. Heckerling, P. L. Fine, T. M. Miller and V. Abraham, "Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation: A Multisite Study of 2 Systems," *JAMA,* vol. 282, pp. 1851-1856, 1999.