Probabilistic Characterization of Neuromuscular Disease: Effects of Class Structure and Aggregation Methods

by

Charles Farkas

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2010

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

Neuromuscular disorders change the underlying structure and function of motor units within a muscle, and are detected using needle electromyography. Currently, inferences about the presence or absence of disease are made subjectively and are largely impression-based. Quantitative electromyography (QEMG) attempts to improve upon the status quo by providing greater levels of precision, objectivity and reproducibility through numeric analysis, however, their results must be transparently presented and explained to be clinically viable.

The probabilistic muscle characterization (PMC) model is ideally suited for a clinical decision support system (CDSS) and has many analogues to the subjective analysis currently used. To improve disease characterization performance globally, a hierarchical classification strategy is developed that accounts for the wide range of MUP feature values present at different levels of involvement (LOI) of a disorder. To improve utility, methods for detecting LOI are considered that balance the accuracy in reporting LOI with its clinical utility. Finally, several aggregation methods that represent commonly used human decision-making strategies are considered and evaluated for their suitability in a CDSS. Four aggregation measures (Average, Bayes, Adjusted Bayes, and WMLO) are evaluated, that offer a compromise between two common decision making paradigms: conservativeness (average) and extremeness (Bayes).

Standard classification methods have high specificity at a cost of poor sensitivity at low levels of disease involvement, but tend to improve with disease progression. The hierarchical model is able to provide a better balance between low-LOI sensitivity and specificity by providing the classifier with more concise definitions of abnormality due to LOI. Furthermore, a method for detecting two discrete levels of disease involvement (low and high) is accomplished with reasonable accuracy. The average aggregation method offers a conservative decision that is preferred when the quality of the evidence is poor or not known, while the more extreme aggregators such as Bayes rule perform optimally when the evidence is accurate, but underperform otherwise due to outlier values that are incorrect.

The methods developed offer several improvements to PMC, by providing a better balance between sensitivity and specificity, through the definition of a clinically useful and accurate measure of LOI, and by understanding conditions for which each of the aggregation measures is better suited. These developments will enhance the quality of decision support offered by QEMG techniques, thus improving the diagnosis, treatment and management of neuromuscular disorders.

Acknowledgements

I would like to thank my supervisor, Dr. Daniel Stashuk, for his valuable guidance and encouragement. Dan's diverse experience and dedication to the success of his graduate students led to many worthwhile discussions that were instrumental in completing this thesis.

I would also like to thank Dr. Tim Doherty and his graduate student, Kendra Derry, for insight and understanding of clinical practice and for providing the clinical data used in this work.

Andrew Hamilton-Wright deserves significant thanks for his implementation of the Pattern Discovery algorithm based on the ideas of Andrew Wong and Yang Wang. The quality of this software greatly improved productivity.

I would like to thank the members of my reading committee, Dr. Catherine Burns and Dr. David Clausi, for their effort and patience in evaluating this thesis.

My deepest gratitude goes to my family and friends for their love and support, while they waited patiently for me to resurface from the piles of research papers on my desk.

Dedication

This thesis is dedicated to my number one grandfather, Charles Juhasz. I am grateful for all the support you have provided over the years, without which I could not have succeeded in my endeavors.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Neuromuscular disorders change the underlying structure and activation of motor units (MUs) within a muscle, thus altering the electrophysiological characteristics of detected electromyographic (EMG) signals. A needle EMG examination is a diagnostic test used to investigate these disease-related changes. Currently, any inference or decision based on the acquired signals is usually accomplished subjectively, and is driven by the expert knowledge of the physician. While there is a small degree of reliance on the most rudimentary forms of quantitative analysis (that is, simply the reporting of numerical figures), diagnosis is usually impression-based and must be made in conjunction with several other diagnostic tests. Quantitative Electromyography (QEMG) is the process of extracting numeric information related to the morphology and activation of MUs from electromyographic (EMG) signals (Stashuk and Brown, 2002). The increased level of precision and reproducibility obtained by using QEMG makes it useful for exploration of the underlying structure and operation of the peripheral neuromuscular control system. Thus, quantitative techniques make it possible to objectively characterize muscle tissue with respect to the presence or absence of disease or distress, and form the basis for clinical decision support.

Recent advances in QEMG techniques have resulted in the quantization of important aspects of needle EMG examinations, resulting in a more precise numerical and statistical representation of their results (Stashuk and Brown, 2002; Stashuk, 1999; Stashuk, 2001; Doherty and Stashuk, 2003; McGill et al., 1991; Pino et al., 2010; Pino, 2009; Pino et al., 2008). The more sophisticated of these methods provide interpreted muscle characterizations that support the decision making process. However, despite the advent of quantitative techniques and the prevalence of decision support in the general field of medicine, QEMG practices remain in their infancy. One of the main impediments to the wide-spread acceptance of QEMG techniques is the inherent difficulty of presenting vast amounts of quantified data in a way that can be easily interpreted. Human decision makers have difficulty interpreting and mentally comparing numeric information without the necessary context or level of explanation. Such shortcomings of quantified data highlight a specific characteristic that decision support and QEMG methods must possess, namely transparency. A system that can explain its conclusions in a manner that is consistent with, and analogous to, the cognitive processes employed by a human decision maker, can thus improve upon the quality and accuracy of any inferences made by the clinician.

The current state of the art in QEMG is the probabilistic muscle characterization (PMC) framework presented in this work and others (Pino et al., 2010; Pino et al., 2008; Pino et al., 2008; Hamilton-Wright et al., 2010; Pino et al., 2008). The key benefit of quantitative methods and PMC in particular, is that they offer greater objectivity and consistency over their subjective counterparts. Providing physicians with objective quantitative information that is supplementary to the information currently obtained from clinical tests can reduce the ambiguity in the diagnosis and treatment of disorders, and provide a higher level of precision and detail. Specifically, measures of confidence or confusion surrounding a particular decision, and information about the types of errors and quality of data make it possible to consider the entire decision making workflow, while providing a transparent explanation of the underlying mechanisms involved in deciding on a course of action. Additionally, the ability to report a richer set of information allows the decision maker to gain insight into the level of involvement (LOI) of disease, allowing clinicians to manage and treat its progression over time. Moreover, the PMC framework, as will be seen, provides a quantitative analogue to current clinical practice. By providing a decision-making framework that mimics the cognitive processes of a human decision maker facilitates easier adoption of, and trust in, such systems by clinicians.

These latter aspects of decision support and QEMG form the fundamental motivation for the current work. In particular, building on the probabilistic muscle characterization framework developed by Pino (Pino, 2009; Pino et al., 2010) and others, the optimal configuration and form of the system is sought. Specific objectives pertaining to these areas include the following:

1. A decision-making (aggregating) scheme is desired that balances the trade-off between sensitivity and specificity.
2. The definition and clinical utility of a measure of LOI is sought.
3. The validity of these methods applied to real-world data is to be evaluated.

A clinically viable decision support system centered on the PMC framework is to be further developed using these principles, and is one of the main goals and motivations for the author and this work.

Overview of the Thesis

The contents of this thesis are dedicated to achieving the above-mentioned goals as follows.

Chapter 2 will provide the reader with the necessary background in neuromuscular disease, clinical electromyography, QEMG, and decision support in order to appreciate the methods described herein.

Chapter 3 will present the probabilistic muscle characterization framework in the context of decision support, and will highlight the key areas that are being further developed in this work. The data sets used in evaluating and validating the methods developed are then presented in Chapter 4, along with a brief statistical analysis in order to build a case for using more advanced techniques.

Chapter 5 describes the methods developed in this work in order to improve the accuracy of disease and LOI characterization, as well as the techniques used in evaluating their performance. Chapter 6 will present the results generated by evaluating the methods, with a focus on those techniques that optimize the abovementioned goals. Following that, Chapter 7 discusses the results in terms of their relevance for clinical use, and highlights any shortcomings of the methods.

Finally, Chapter 8 concludes the work by summarizing the relevant findings and assessing whether or not they meet the goals outlined in the introduction. A special sub-section will make note of recommended future work, that the author feels is necessary in order to adopt the developed system for clinical use.

Appendix A provides supplementary material pertaining to derivations of certain equations or mathematical background material that is not well-suited for the background section of this thesis.

Appendix B presents screen-shots of a decision support system implementation based on the probabilistic muscle characterization framework.

Chapter 2
Background

2.1 Physiology and Neuromuscular Disease

The smallest functional unit of muscle is termed the motor unit (MU), and consists of a group of localized muscle fibers (MFs) that are innervated by a single neuron in the spinal cord. MUs are recruited (activated) by the central nervous system according to Henneman's Size Principle[1] and it is the combination of the firing frequency of each MU, as well as the number and size of active MUs, that determine the amount of force exerted by the muscle. Contraction of muscle fibers is preceded by the initiation of action potentials in the end plate region and subsequent spread of the potentials along the surface of the sarcolemmal membranes toward the musculo-tendinous junctions of the muscle fibers. The currents associated with these action potentials spread throughout the extracellular space surrounding the muscle fibers and create time varying voltage fields (Stashuk and Brown, 2002).

The latter field potentials can be detected by needle electrodes positioned in the extracellular space and the resulting waveform measurements form the basis of needle electromyography. Waveforms generated by action potentials of single muscle fibers (MF) are called muscle fiber potentials (MFPs) whereas waveforms associated with the electrical activity of whole motor units (MU) are called motor unit potentials (MUPs). The successive, semi-rhythmic MUPs generated by an active MU are referred to as motor unit potential trains (MUPT) and are often represented by a single aggregate MUP template with similar characteristics. The collection of waveforms detected during the contraction of a muscle is termed an electromyographic (EMG) signal and represents the spatial and temporal summation of all contributing MUPTs as well as baseline noise.

Neuromuscular disorders (NMD) change the morphology and activation patterns of the MUs of the muscles affected. There are two major categories of NMD: those that affect the nervous system are termed neurogenic (e.g., Lou Gehrig's disease), while those that affect the muscle tissue are referred to as myopathic (e.g., Muscular Dystrophy). As depicted in Figure 2.1, myopathic disorders in general maintain the number of MUs in a muscle but reduce the number of muscle fibers in the MUs

[1] Henneman's size principle states that motor units are generally recruited in order of smallest to largest (fewest fibers to most fibers) as contraction increases. An electromyographer must be aware of this when sampling MUs because inconsistent contraction levels may introduce abnormally large or small MUPs resulting in possible data misinterpretation.

due to atrophy and eventual fiber necrosis. At the same time, other fibers may hypertrophy thus increasing the range of fiber diameters and motor unit fiber densities in the MUs. To compensate for fiber loss and fiber diameter changes, MU recruitment and firing rate are higher in myopathic muscle then in healthy muscle in order to create a specific level of contraction. In contrast, neurogenic disorders in general cause a loss of motoneurons or innervation, resulting in fewer MUs. However, denervated fibers are reinnervated by axonal sprouts of surviving MUs resulting in muscles with fewer MUs but with MUs that have larger territories, larger numbers of fibers, and uneven motor unit fiber densities compared to healthy muscles. These MU fiber changes are also depicted in Figure 2.1. As a result of MU loss and reinnervation, fewer MUs are required to be recruited in neurogenic muscle than in healthy muscle to create a specific level of contraction, but fine motor control is diminished.

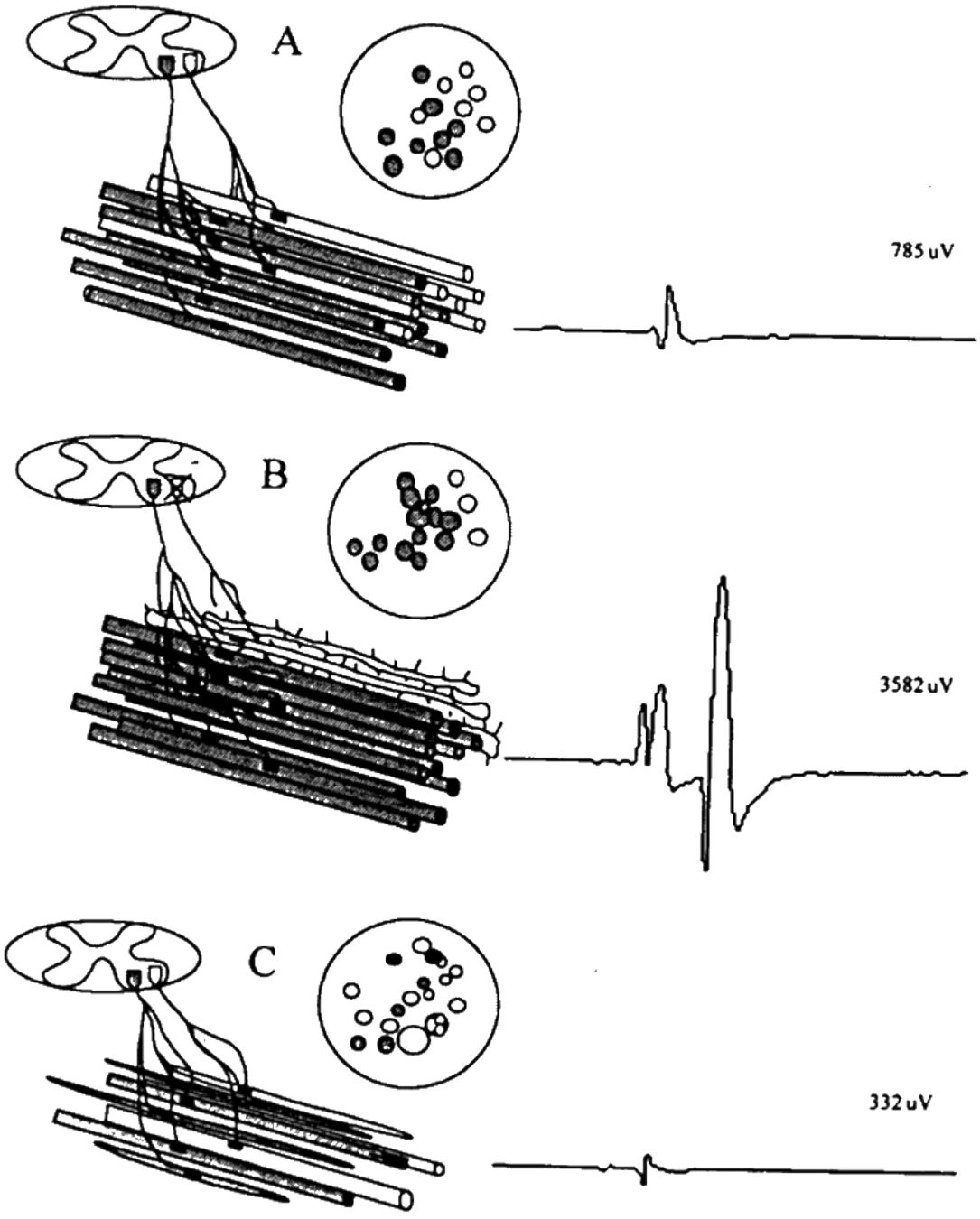

Figure 2.1: Effects of disease on motor unit morphology and motor unit potentials

2.2 Clinical Electromyography

EMG signals provide a unique insight into both the structure and function of muscle tissue. The fiber and MU activation changes caused by a disorder are reflected in the characteristics of needle EMG signals allowing them to be used to help diagnose treat and manage neuromuscular disorders. Currently, physicians interpret the findings of needle EMG examinations qualitatively, and are often prone to bias. There is some reliance on quantitative methods; however these methods are somewhat rudimentary. For instance, formal quantization was introduced by Buchthal in the late 1950's, wherein normative limits of common MUP features were established. Since then many further attempts have been made to establish predictive values and normative limits for specific muscle groups, as seen in the works of Podnar (Podnar and Vodusek, 2001; Podnar and Mrkaic, 2002; Podnar, 2009b; Podnar, 2009a; Podnar, 2004b). Quantitative analysis is able to provide better accuracy of disease detection in a way that is more reliable and objective, but suffer other limitations. In particular, the vast amounts of numeric data are difficult to interpret, creating many challenges in explaining and presenting such findings in a concrete way.

2.2.1 Conventional Needle Examination

Current clinical practice is to qualitatively interpret the findings of needle EMG examinations to infer the presence or absence, as well as the type of a neuromuscular disorder. Although subjective, a qualitative assessment allows an experienced physician to infer a host of underlying conditions and diseases by assessing insertional, spontaneous and voluntary needle EMG activity. While lacking the precision of quantitative analysis, qualitatively extracted information can provide evidence related to disease categories, as well as specific disease processes.

During a conventional needle EMG examination (Daube and Rubin, 2009), a needle electrode (monopolar or concentric) is inserted into the superficial layers of a muscle and the resultant detected EMG signals are assessed based on three types of recorded activity: insertional, spontaneous, and voluntary. In all cases, the electrode must be positioned to sample several regions of the muscle to obtain a sufficient statistical sampling of the MUs. When assessing voluntary activity, EMG signals are assessed at a low level of contraction with only a few active motor units contributing to the signal. However, if decomposition techniques are used EMG signals with 5-7 MU contributions can be assessed. Either way, it is important to position the needle to detect MUPs with a rapid rise time to ensure that the detection surface is sufficiently close to the fibers of the active MUs. It is also

important to ensure that the level of each contraction is consistent, so that abnormally large or small MUPs may be attributed to morphological changes in the muscle and thus representative of disease and not be related to different sizes of active MUs due to differing levels of muscle activation. QEMG techniques have not as of yet been applied to study EMG signals related to insertional or spontaneous muscle activity. As such the following will focus on EMG signals detected during voluntary muscle activity.

2.2.2 Qualitative Characterization

Traditionally, needle EMG signals detected during voluntary activation are qualitatively characterized based on a visual and auditory assessment of the morphology, stability, and times of occurrence of their MUPs. In addition, the intensity of composite EMG signals or interference patterns are described using terms ranging from full to sparse in an effort to assess MU activation. These characterizations are usually manually quantized and charted. While such qualitative assessment is prone to a high error rate, a skilled practitioner can use it to detect not just the broad category of disease, but also certain processes that are symptomatic of a specific disease or group of disease processes.

MUP analysis is useful in determining the type of disorder, (myopathic or neurogenic) as well as the time course and severity of the disease. However, qualitative analysis is limited to the MUPs of only a few active MUs at a time. MU morphology is inferred based on assessment of the duration, amplitude and number of phases of detected MUPs. Duration is related to the number of MU fibers and the size of the MU territory. Amplitude reflects the contribution of fibers that are nearest to the electrode, and can provide an indication of the number and diameter of MU fibers. Amplitude and duration typically increases in neurogenic disease and decreases in myopathy. The number of phases refers to the number of baseline crossings of a MUP. It is a non-specific measure of complexity that is related to the synchrony of muscle fiber activity and can indicate the presence of a neurogenic or myopathic disorder. MUP stability refers to consistency of MUP morphology across the MUPs of a MUPT. Unstable MUPs are a sign of impaired transmission across the neuromuscular junctions (NMJs) of a MU. Although MUP instability is typically a sign of a disease of the NMJ, such as myasthenia gravis, any disorder associated with denervation/reinnervation may cause unstable MUPs. During qualitative assessment, a physician keeps track of abnormally small or large MUPs and polyphasic MUPs and assigns rankings based on their severity and frequency. Once a sufficient

number of samples are obtained, the physician forms a subjective impression of the underlying disease process based on the evidence obtained.

MU activation patterns are assessed in terms of the level of MU activation and recruitment. Low activation represents a central process such as a CNS disorder or manifestation of pain. Compromised MU recruitment is found in neurogenic diseases, and sometimes in end-stage myopathy, whereas early recruitment (i.e., increased activation) is typically a sign of myopathy. According to Shapiro et al. (Preston and Shapiro, 2002), MU activation analysis is one of the most difficult tasks for an electromyographer. Qualitative MU activation analysis is limited to very low levels of contraction where only a small number of MUs are active. Once a sufficient number of MUs are recruited, the overlap of MUPs produces an interference pattern and MU activation becomes qualitatively indiscernible. Daube (Daube and Rubin, 2009) and Shapiro (Preston and Shapiro, 2002) describe the normal and abnormal variations of MU activation characteristics, and provide further references to standardized values for manual visual and auditory assessment.

2.2.3 Quantitative Electromyography

QEMG techniques attempt to reproducibly extract useful muscle information from sets of features values used to represent important EMG signal characteristics. Features can represent characteristics of a composite EMG signal (i.e., interference pattern) or, of individual motor unit potential trains (MUPTs) extracted from a composite EMG signal using either simple triggering techniques or more sophisticated decomposition techniques. With respect to the latter, time domain features related to MUP size, shape and stability or features of transformed or modeled MUP waveforms have been used.

Features Used in Quantitative Analysis

In general, the most immediate benefit of quantitative analysis is the provision of more precise continuous values for features like duration, amplitude and number of phases/turns. Continuous-valued features also make it possible to derive relationships among features, such as area, thickness and size index (Sonoo, 2002), as well as measures of irregularity, such as the 'irregularity coefficient' (Zalewska and Petrusewicz, 2005; Zalewska et al., 2004). With the help of Figure 2.2, the commonly-used features that describe MUP size, shape and complexity are defined next.

Duration (ms):	The time between the starting (onset) and end point of a MUP. These points are often determined using deviation from the baseline and MUP slope criteria.
Amplitude (uV):	The difference in voltage from the maximal negative to maximal positive peak within the duration of a MUP.
Area (uV.ms):	The summation of the rectified MUP signal within the duration.
Thickness (ms):	The area-to-amplitude ratio (AAR) (Podnar, 2009b).
Size Index:	A logarithmic function of thickness and amplitude that is related to the size and shape of a MUP (Sonoo, 2002).
Number of Phases:	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.
Number of Turns:	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.

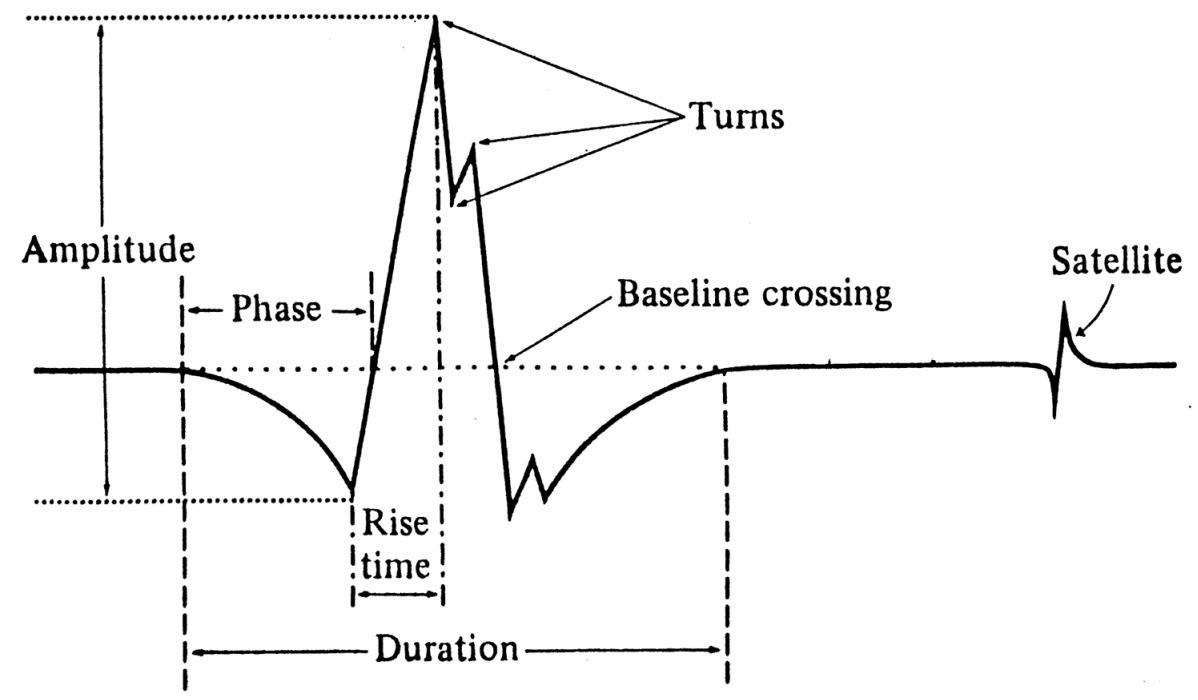

Figure 2.2: Features of a Motor Unit Potential (From Stashuk and Doherty: "Normal Motor Unit Action Potential" in Neuromuscular Function and Disease, vol. 1, Brown, Bolton and Aminoff, Eds. Philadelphia, PA: Elsevier Science, 2002, pp. 291-310.)

2.2.4 Quantitative Characterization

QEMG methods use rigorous statistical and/or probabilistic inference, where feature values from EMG signals detected in the muscle under examination (i.e., test EMG signals) are compared with respect to distributions of corresponding values obtained from exemplary EMG data. By describing the statistical similarities or differences between test values and exemplary data, it is possible to characterize EMG signals, or constituent MUP waveforms acquired from a muscle under examination. This allows a clinical decision maker to use, to the fullest, the information present in the test EMG signals, and in addition, to leverage information extracted from previously acquired (exemplary) signals. Such exemplary data must be acquired and stratified by an expert clinician. In addition, it must be specifically acquired for each muscle or groups of muscles studied, and should be stratified by age or any other relevant stratification criteria (e.g., gender) and incorporated into a repository (database).

QEMG analysis has the advantage of providing greater objectivity and consistency, and is useful for equivocal cases to increase the certainty of a diagnosis. Furthermore, the precision obtained with more sophisticated quantitative techniques can provide continuous measures that also relate to the level of involvement (LOI) of a disorder (Pino and Stashuk, 2008). Compared to qualitative methods QEMG methods can be more robust and reproducible. They can provide analytic confidence (i.e., provide measures of uncertainty or error) and have the ability to generalize across sets of EMG signals.

The reproducibility and robustness of EMG data assessment are two important criteria used to evaluate QEMG methods. Being able to discuss the degree to which similar features values can be obtained from similar EMG signals, coupled with discussions related to the sources and types of confounding error present, makes it possible to consider the entire workflow of obtaining decision-making data from an EMG signal in terms of reproducibility, robustness, and in general the data quality associated with the outcome measures associated with EMG based analysis. Whether EMG data is used for exploration of muscle structure and function or to detect and characterize disease it is clear that metrics for the quality of the data are invaluable, and by measuring and improving data quality, any inference made based on this data will also be improved.

Regardless of whether a qualitative or quantitative assessment is used a similar hierarchical process is followed. First, in order to account for the large variability in MU size and MUP shape throughout a muscle, MUP or/and other signal features are assessed at several needle positions within a muscle. The data from these various needle positions are then characterized based on whether they possess attributes consistent with certain disease processes. The characterized sampled data are then combined to arrive at an overall impression or characterization of the muscle. Finally, a rule or heuristic is applied to categorize the muscle based on the characterization measures obtained.

2.3 Decision Support Systems

One of the main impediments to the wide-spread acceptance of QEMG techniques is the inherent difficulty of presenting vast amounts of quantified data in a way that can be easily interpreted. Human decision makers have difficulty to interpreting and mentally comparing numeric information without the necessary context or level of explanation. Such shortcomings can be overcome by augmenting certain QEMG methods within a clinical decision support framework that transforms the vast amounts of quantitative data into clinically useful muscle characterizations.

Numerous methods of analyzing and interpreting data to provide decision support have been reported in the literature. This is especially the case in medicine where large amounts of disparate data are analyzed and amalgamated to make critical and time-sensitive decisions. One of the most important aspects of a clinical decision support system is its ability to combine two important (and sometimes conflicting) capabilities: the ability to characterize consistently and accurately make a prediction, and the ability to explain that prediction in a manner that is easily understood by the decision maker.

Therefore, one of the main criteria used in evaluating a CDSS is transparency. Transparency is especially important when dealing with complex data structures, such as multivariate distributions of MUP feature values or other statistical parameters. A system that can explain its conclusions in a manner that is consistent with, and analogous to, the cognitive processes employed by a human decision maker, can thus improve upon the quality and accuracy of the inferences made. Therefore, the degree to which the findings and supporting evidence can be presented and explained to support the decision making process is an important factor to consider when developing a clinical decision support system. How the findings and supporting evidence are presented and explained to support the decision making process is also very important.

2.3.1 Decision Support in Medicine

The prevalence of clinical decision support systems in medicine is growing exponentially, and systems have been proposed, both in the field of neuromuscular disease, and in diagnostic medicine in general. For example, the MUNIN system (Suojanen et al., 2001) uses Bayesian Networks to infer a disease based on findings from clinical tests, examinations and patient reports. A major limitation of this system is that a human medical expert is required to 'train' the system by defining conditional probability tables. A machine that can 'learn' without a great deal of user intervention would be an asset. The KANDID system (Fuglsang-Frederiksen et al., 1993) uses logical operators and known rules to build a knowledge base. The system is used by making queries that are answered using 1st order logic.

In the field of medicine in general, decision support systems are being developed at a rapid rate. In most cases, data from multiple sources, such as clinical evaluations, imaging, and other medical tests, are used to achieve a more robust inference. For example, miniTUBA (Xiang et al., 2007) is a medical inference system based on dynamic Bayesian networks (DBN). DBNs are able to interpret

heterogeneous, fluctuating data and are able to capture time varying clinical parameters as well as predict the course of disease progression. While for breast cancer diagnosis (Revett et al., 2005) developed a system by combining 'rough sets' and neural networks and (Gevaert et al., 2006) developed a breast cancer prediction scheme based on Bayesian networks which combines high-dimensional microarray data with clinical findings. In the field of pulmonary diseases, Economou, Goumas and Spiropoulos (Economou et al., 1996) initially used ANNs in a decision support system and then developed a knowledge base modeled after the methodology used by physicians for clinical differential diagnosis to provide decision support (Economou et al., 2001).

2.3.2 Requirements for CDSS

Given this widespread use of decision support, attempts have been made to formalize the general requirements for clinical decision support in order to guide the development and evaluation of future systems. The works of Konokenko and Sprogar (Kononenko, 2001; Šprogar et al., 2002) and a later summary by Pino (Pino, 2009) have generated a list of requirements for clinical decision support methods to be clinically useful and safe. These are:

Transparency:	The system must be able to intuitively explain the mechanisms of inference, findings, and supporting evidence;
Accuracy:	Performance of the system must be superior to conventional methods;
Confidence:	A measure of confidence or certainty in the inference or assertion made by the system must be provided. Alternatively, the system should fail gracefully when a decision cannot be made with certainty, as opposed to reporting an incorrect result with high confidence;
Numeric Value:	It is desirable to have a numeric or continuous characterization value or measure to facilitate comparisons and longitudinal studies;
Mixed-Mode Data:	The system should be able to handle all types of input data, including continuous, discrete, and categorical data;
Multivariate:	Multiple inputs or features must be considered simultaneously so that higher order relationships and patterns can be observed;

Generalization: The system must be able to make accurate predictions/decisions when presented with novel input patterns;

Missing Data: While missing input values may or may not affect the confidence in a decision, it must not compromise the system's ability to function.

In addition to these requirements, it is desirable to have a reasoning strategy and flow of information that, even in an abstract way, is analogous to the cognitive processes and data assimilation strategies employed by a human decision maker. This helps to establish trust in the system, and provides a 'consultative' approach, rather than a simple statement of the facts.

2.3.3 Decision Theory and Evidence Aggregation

Most decision support systems have a similar hierarchy or structure. At the lowest level, large amounts of data (usually numeric) are extracted from one or more clinical test or procedure. These data are then combined to support a particular action or decision. The evidence can come from disparate clinical tests, as is the case in (Xie et al., 2005), where the results from each test can have a different form and must be combined using a data fusion approach. Alternatively, the evidence can represent a single test repeated multiple times, as is the case with QEMG examinations, where a statistical sampling of MUP morphology is obtained from various needle positions throughout the muscle.

The reliability and diagnostic yield of this data can often vary among tests, as well as within a particular procedure. Thus the aggregation step plays a critical role in formulating an overall decision. As such, when discussing the transparency of decision support, it is important to consider both the transparency of the underlying classifier as well as how the information is structured to support the decision making process. Essentially, the flow of information from low-level evidence to high-level actions or suggestions must be presented to the clinician in an intuitive way.

The hierarchical muscle characterization process described previously follows this model quite closely. Each MUP characterization is essentially a piece of evidence, provided in the form of a conditional probability vector. Each piece of evidence is then aggregated by the CDSS in order to make an overall decision about the muscle. By forming a decision in this way, it is possible to 'drill down' into each piece of evidence, to determine how it was characterized as well as how it was factored into the outcome (e.g., how significantly it was weighted or whether it was a source of

confusion). The similarity between this model and the cognitive decision making process employed during qualitative interpretation of needle EMG findings provides an effective way to distil and communicate the findings of quantitative analysis.

2.3.3.1 Decision Theory

The strategies employed in arriving at a decision depend on the amount and quality of evidence available. Decision makers often follow two main paradigms when evaluating information, depending on the quality or confidence in each piece of evidence. The theories of decision making and analysis thus play a crucial role in formulating the aggregation stage of a CDSS, and different strategies can be employed depending on the objectives of the decision being made. Such objectives might be to maximize characterization accuracy, balance trade-offs between different types of errors (type I vs. type II errors) (Duda et al., 2000), or to obtain accurate estimates of decision confidence. A brief discussion of relevant aspects of decision theory is thus required.

Decision analysis is the process of structuring and decomposing hard decision problems into their key components and identifying the best option through a systematic approach (Budescu and Yu, 2006; Clemen and Reilly, 2000). A common situation arises in decision making where the decision maker has access to multiple estimates of the probability of an important target event, which need to be combined into a single estimate of the certainty in the resulting decision. There are several reasons why a decision maker would want to combine multiple sources of information, namely to maximize the amount of information that the decision is based on, reduce the influence of extreme sources (i.e., outliers), and to reduce the effect of unreliable or inaccurate information (Budescu and Yu, 2006). Several studies have shown that combining such pieces of evidence greatly improves the quality of prediction of QEMG techniques, and in general (Pino et al., 2009; Pfeiffer and Kunze, 1995; Pfeiffer, 1999; Hamilton-Wright et al., 2010; Hamilton-Wright and Stashuk, 2006).

Decision making strategies can be based on two classes of models, namely simple averaging and Bayesian (naïve) analysis. While both models have similar accuracy in terms of making the correct decision, their underlying principles are different leading to radically different estimates in the confidence in a decision. For instance, the averaging model weights each piece of evidence as an inverse function of the total number of items, reducing the effect of each piece of information as more evidence is made available. In contrast, the naïve Bayes' model promotes an additive effect where each piece of evidence contributes to the decision without accounting for the total amount of

information. Essentially, the averaging rule produces a compromise effect while the naïve Bayes' rule produces an extremeness effect (Budescu and Yu, 2006).

As mentioned, an important aspect of decision making is the confidence in the decision being made. Budescu states that the most important predictors of confidence are the level of agreement or disagreement between the pieces of evidence, the estimated predictability or likelihood of an event, and the total amount of information available. Based on these predictors, it is possible to develop aggregation techniques that are a compromise between the two extreme models of averaging and Bayesian aggregation. Budescu develops two such models, the weighted mean log-odds (WMLO) and the Adjusted Bayes (AB) rule. The WMLO rule is based on averaging, but accentuates the differences for extreme values, stretching them out prior to averaging. The Adjusted Bayes rule has the effect of pulling in extreme values prior to applying Bayes' rule. Since these measures capture the important aspects of both extremes of reasoning, it is expected that they will perform better on average under random variations of the above factors.

Chapter 3
Muscle Characterization and Clinical Decision Support

Given the wide variability of MUPs detected in a muscle dependent on the specific needle positions during signal detection and the varying degrees to which a disorder may affect specific MUs, individual MUP characterization scores are not robust indicators of the actual category or state of a muscle (i.e., myopathic, normal or neurogenic) and cannot be used in isolation for accurate diagnosis. A more robust indicator can be achieved if several MUP characterizations across a muscle are aggregated. During qualitative analysis a clinician also aggregates, based on experience and training, the information extracted from the MUPs and interference patterns examined to create an overall clinical impression of the muscle under examination. Unfortunately, the consistency and accuracy of the clinical impression obtained is dependent on the experience of the clinician. However the decision support structure and evidence aggregation methodologies presented in the previous section are conducive to quantitative muscle characterization based on a statistical sampling of MUP characterizations.

Muscle characterization for decision support is accomplished in two major ways, with varying degrees of success in meeting the requirements laid out previously. The simpler and more commonly used method relies on statistical analysis of average MUP feature values relative to exemplary data. Formal standards and criteria have been laid out to guide this process, and basic decision support can be provided through the reporting of such statistics. Inference based on statistical data is commonly achieved by applying rules or heuristics to certain statistical properties, or may be entirely left up to the clinician. Probabilistic muscle characterization, on the other hand, provides a more transparent and intuitive way to aggregate MUP characterizations, forming an overall muscle characterization. This method is able to numerically quantify the degree to which a disease is present, along with a measure of confidence in its estimate, and is based on a multivariate feature analysis.

3.1 Statistical Muscle Characterization

The most common method to address the shortcomings of subjective qualitative interpretation is to use summary statistics of quantified MUP features. Statistical techniques go beyond simply summarizing quantitative data, and attempt to characterize a muscle based on the distributions of their

sampled MUP parameters. Such predictive analysis techniques form the basis of the early decision support techniques.

Common practice has been to calculate mean values for morphological MUP features, such as duration, amplitude and number of phases across sets of MUPs and compare them to normative standards reported in the literature to get a muscle-level impression of the state of disease involvement. For example, the early works of Buchthal (Buchthal et al., 1954) are still used as standards for formal quantization, and Podnar (Podnar, 2009a; Podnar, 2009b; Podnar and Mrkaic, 2002) has published predictive values for limb, genioglossus, and anal sphincter muscles. Researchers have cited constraints on the minimum number of MUs sampled in a muscle (i.e. the preferred number of MUPs sampled for robust diagnosis) and the general consensus is that a minimum of 20 MUPs should be sampled (Podnar and Mrkaić, 2003). However, Podnar showed that sampling additional MUPs can increase sensitivity and specificity (Podnar, 2004a). The major drawback with the statistical methods is that for each MUP feature suitable threshold values for defining normality/abnormality must be established as well as rules regarding the number of abnormal features required to declare a muscle as abnormal.

To implement a statistical method, a training data set consisting of exemplary 'normal' MUPs is used to generate distributions for each feature. Mean (Stewart et al., 1989) and/or outlier (Stålberg et al., 1991) threshold values can then be defined, and feature values of a muscle under test are classified based on their locations relative to these limits. For example, Stalberg used a mean normative range of +/- 2 standard deviations and considered three different outlier criteria: 95% confidence limits; 95% confidence limits for the third highest and third smallest values; and extreme upper and lower outlier limits. No appreciable difference was found across these methods. A further strategy based on this approach applies rules and heuristics to interpret the numbers of outliers and deviations from normal mean limits (Pino et al., 2008; Pino et al., 2010; Pino and Stashuk, 2008). Several works by Podnar (Podnar, 2004b; Podnar, 2005; Podnar, 2008) have performed comparisons of different outlier criteria and have established standardized normative limits for diagnostic criteria.

The intention of QEMG analysis is to increase both sensitivity (number of correct positive decisions) and specificity (number of correct negative decisions). For every decision, the objective is to maximize sensitivity and specificity. However, sensitivity and specificity cannot be simultaneously maximized and often involve some amount of trade-off or compromise. Therefore, with regard to the

decision making process, whether it is better to be less specific in order to be more sensitive must first be made. Best practice is to determine the minimum number of overall errors by balancing sensitivity and specificity unless the cost of poor sensitivity relative to specificity or vice versa can be determined. In this regard, statistical muscle characterization methods have two main implementation issues: First, it is difficult to suitably define thresholds of normality/abnormality. Second, much confusion exists in determining suitable categorization rules. For instance, a rule might state that a minimum number of outliers must be present to declare abnormality, but results may vary depending on which features happen to have outliers. On the other hand, a single feature might have sufficient evidence due to the number of outlying MUPs. Such a feature may or may not fail the 'mean' test depending on the distribution of its MUP feature values but based on outlier criteria clearly demonstrate abnormality.

3.2 Probabilistic Muscle Characterization

Statistical muscle characterizations have two main weaknesses in addition to the implementation issues cited above: 1) they are unable to provide a measure of, the quality of, or confidence in, a particular inference, and 2) they use thresholds defined using averages and deviations across sets of MUPs (20 or more) making it difficult to investigate the rationale for a particular result at the level of the individual MUPs. Both of these weaknesses reduce the transparency of decisions made. A probabilistic approach addresses these weaknesses by providing information at the muscle level as well as the MUP level, thus allowing for the provision of a detailed explanation of the underlying decision-making process. The term 'characterization' in this sense takes on a more complex definition, but with the same end result as described for statistical characterization. A probabilistic characterization is one that assigns a score or likelihood measure to each muscle category under consideration. Ideally the score represents the probability of the examined muscle being affected by a disease of a particular category conditioned by the specific characteristics of the set of its sampled MUPs. A characterization is a set of n scores, where n is number of muscle categories under consideration (typically 2 or 3).

The probabilistic muscle characterization (PMC) framework was first introduced by Pfieffer (Pfeiffer and Kunze, 1995; Pfeiffer, 1999) and further developed by Pino and Hamilton-Wright (Pino, 2009; Pino et al., 2010; Hamilton-Wright and Stashuk, 2006; Pino et al., 2008; Hamilton-Wright et al., 2010). Many studies have shown it to be superior to both statistical muscle characterizations as

well as conventional subjective examinations (Pfeiffer and Kunze, 1995; Pfeiffer, 1999; Stewart et al., 1989; Stålberg et al., 1991; Podnar, 2004b; Podnar, 2005; Katsis et al., 2007; Pino et al., 2008; Pino et al., 2010). The PMC framework relies on a knowledge base of exemplary training data collected from patients with known pathology and healthy controls (see Figure 3.1). The characterization of a muscle under test then follows a process that is similar to a conventional needle EMG assessment of MUP morphology. MUPs are sampled uniformly at points throughout the muscle to obtain a statistical representation. The detected EMG is then decomposed into its constituent MUPs using decomposition techniques described elsewhere (Stashuk, 1999; Stashuk, 2001). The MUPTs are then individually characterized in terms of the exemplary data based on morphological features of MUP shape, size and complexity. The individual MUPT characterizations are then combined into a muscle characterization using a suitable aggregation scheme (see Figure 3.2). A heuristic is then applied to a muscle characterization in order to infer a particular category of disease. In most cases, the standard rule is to select the category with the highest characterization score as the winner. Based on the muscle characterization, a numerical measure of the confidence in the predicted outcome is also calculated. The level of involvement can be obtained in a similar fashion; however appropriate aggregation schemes and techniques have not yet been assessed. One of the goals of this work is to determine if and how the type of aggregation method used affects the way in which characterization scores are calculated, and whether various strategies of compromise or extremism may be more suited to a specific goal. Thus, optimal strategies to maximize accuracy in detecting disease while balancing the amount of risk, along with characterization scores that correlate well with LOI, are sought.

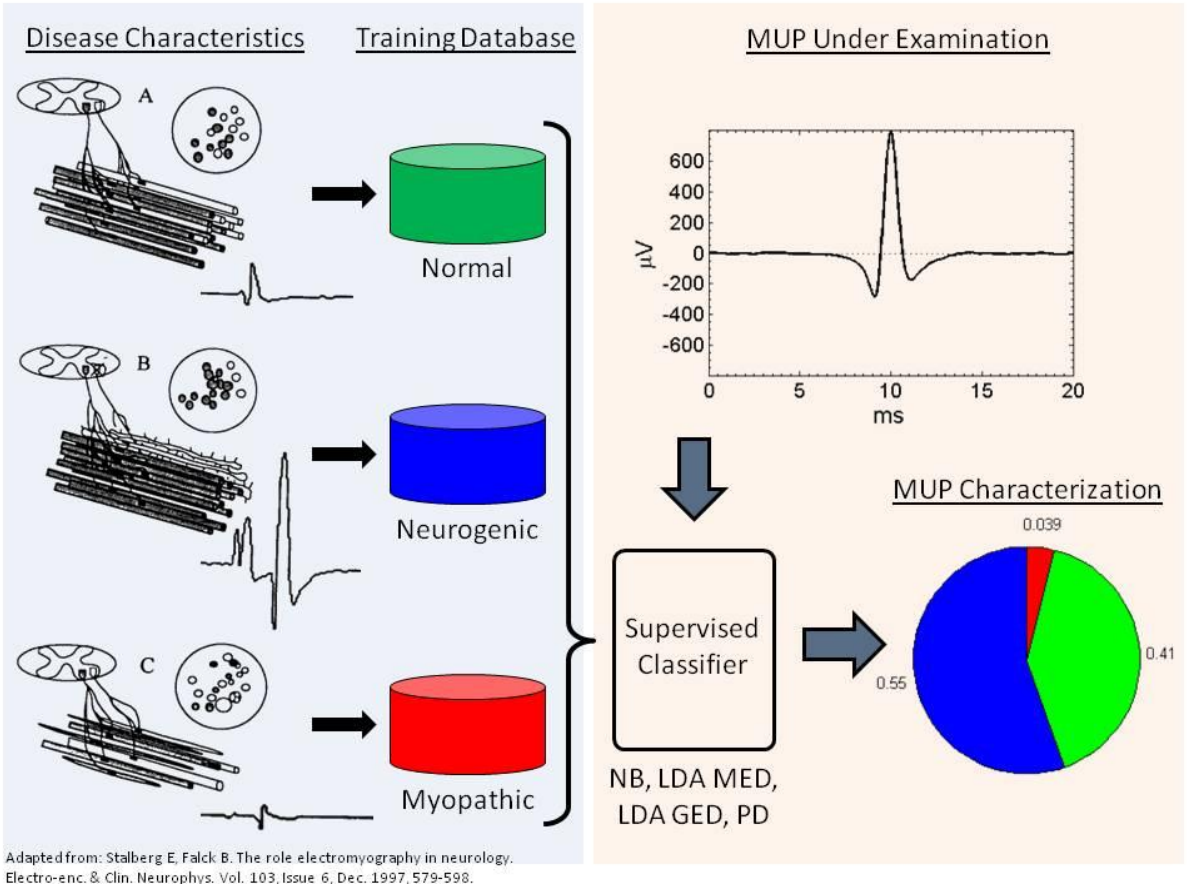

Figure 3.1: Characterization of a MUP under examination based on exemplary data

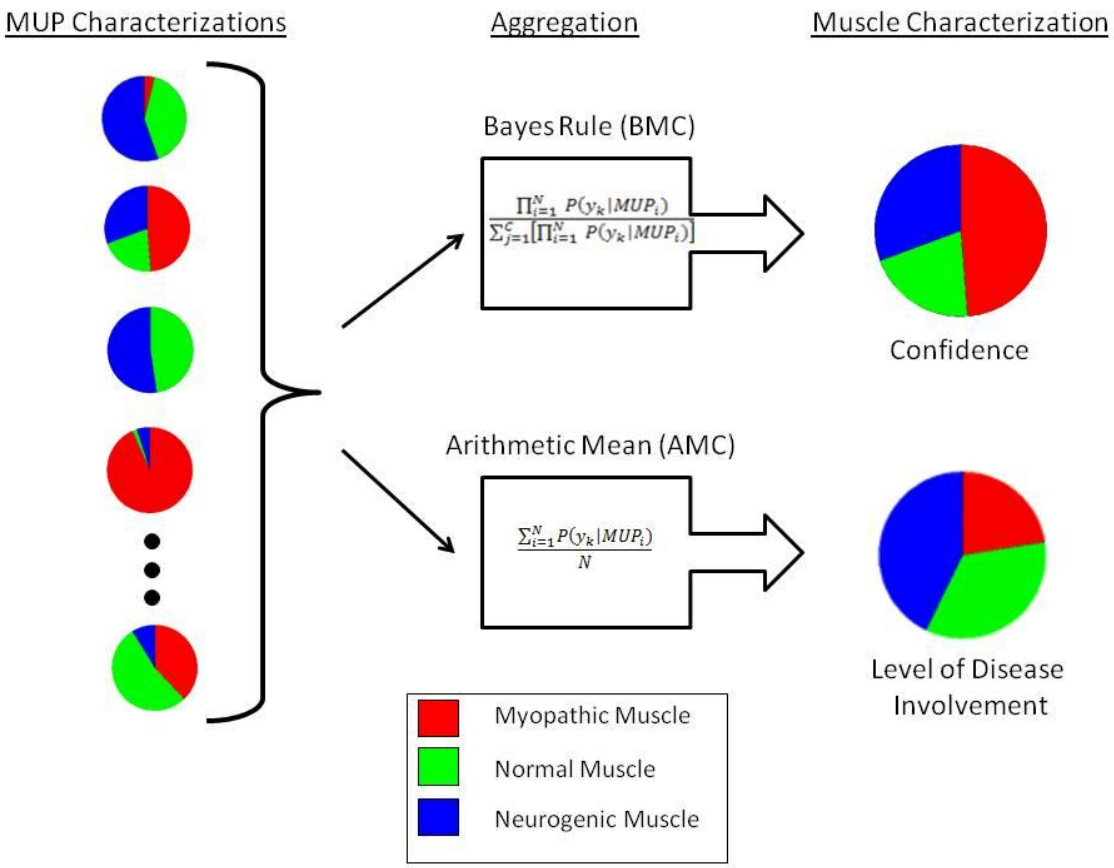

Figure 3.2: Aggregation of MUP scores into a muscle characterization

3.2.1 MUP Characterization

Based on a training set of features extracted from exemplary MUP templates, a MUP characterization consists of a set of scores that indicate the likelihood that the muscle from which a MUP template was calculated is affected by a specific category of disease based on the characteristics of the MUP template. In a two-category case, the likelihood of a muscle being normal or abnormal conditioned on the characteristics of the MUP template is estimated, or in a three category case, which is more common, the likelihood of a muscle being normal, myopathic or neurogenic conditioned on the characteristics of the MUP template is estimated. Alternate class stratification may also be used, as is the case when working with data that represents different levels of involvement, but require training data to be specifically acquired and stratified by an expert.

MUP characterization can be performed using a variety of pattern recognition techniques. Many of the conventional continuous classifiers based on probability theory perform reasonable well; while more advanced techniques have the potential to improve upon one or more of the CDSS requirements laid out previously.

Conventional Classifiers

Several MUP characterization techniques that are based on classifiers that use continuous valued features have been considered in the literature. The most common of these is linear discriminant analysis (LDA) (Duda et al., 2000), and was used by Pfeiffer and Kunze, and later by Pino, Stashuk, and Podnar as a candidate for a muscle characterization methodology (Pfeiffer and Kunze, 1995; Pfeiffer, 1999; Pino, 2009; Pino, et al. 2010; Pino et al., 2008).

LDA statistically transforms data in an attempt to minimize within-class scatter (variance) while simultaneously maximizing between-class scatter to achieve an optimal decision function for each muscle category. Pino and Stashuk (Pino et al., 2008) evaluated the performance of LDA, decision trees (DTs) and a standard Naive Bayes (NB) classifier (Duda et al., 2000), and found that MUP categorization performance was comparable across these methods. The advantage of conventional classifier techniques is that their simplistic nature allows for some degree of transparency, provided that a clinician has a moderate statistical background, although their performance is subject to the data upholding certain assumptions such as obeying a Gaussian distribution and independence across features.

Advanced Pattern Recognition Techniques

When MUP categorization accuracy is considered as the evaluation criteria, conventional classification techniques do not provide the best performance. Therefore advanced pattern recognition techniques have been used to improve performance. In these studies, the true categories of MUP templates were manually determined by an expert neurophysiologist to create a MUP database. In their early works, Pattichis (Schizas et al., 1999; Pattichis, 1999) was the first to develop a decision making strategy based on artificial neural networks (ANNs) combined with Kohonen self-organizing feature maps (SOFM) and a learning vector quantization (LVQ) technique to perform MUP characterization based on morphological MUP features. Katsis (Katsis et al., 2007) used ANNs with radial basis functions and probabilistic neural networks (PNNs) in a two-stage classification

approach. The first stage used an ANN or PNN to discriminate between normal and abnormal MUPs, and the second stage used a C4.5 decision tree (Salzberg, 1994) to determine whether the abnormality was myopathic or neurogenic. Although these advanced pattern recognition based MUP characterization techniques are robust and accurate, several drawbacks have been noted. For example, ANN methods have a tendency toward over-fitting resulting in a difficulty to generalize to new data. In addition, while most of the advanced pattern recognition techniques, including ANNs, SVMs and others, offer greater sensitivity and specificity they do not provide good transparency. It is difficult to intuitively appreciate how they make their decisions. Poor transparency leads to poor acceptance of results by clinicians.

Transparent Rule-based Techniques

The need for transparency can be addressed by employing rule-based methods. The two-stage method developed by Katsis is able to provide a level of partial transparency and interpretability by using a decision tree to discriminate between myopathic and neurogenic categories. However, the most promising techniques involve the use of pattern discovery (PD), developed by Wang and Wong (Wong and Wang, 1997). Pino (Pino, 2009; Pino et al., 2008) demonstrated that PD had comparable performance to the continuous classifier methods while meeting the necessary conditions for clinical decision support.

PD quantizes all feature values into events. In a training set, patterns of events across the features used to represent a MUP and including a specific muscle category, which occur more often than expected under the assumption of independent features and categories, are selected by PD as rules. The entire set of discovered rules form a knowledge base that can be used for categorization of test MUPs. If a pattern of events created by the quantized feature values of a test MUP match a rule associated to a pattern discovered in the training data that occurred more often than expected it is used as positive evidence and supports the category contained in the rule. Alternatively, if the matched rule is associated to a pattern discovered in the training data that occurred less often than expected it is used as negative evidence and refutes the category contained in the rule. The degree of support or refutation is measure by the weight of evidence statistic. The order of a pattern or rule refers to the number of features involved in the rule, including the class label. While PD was been shown to have accuracy comparable to other classification techniques, it may not be considered to be as robust as some of the other classification methods, due to the need to quantize feature values. The number of

intervals ('bins') used for quantization must be carefully selected to manage the trade-off between the level of granularity (and accuracy), and the quality of evidence. For example, if the number of intervals increases, more training data is needed to determine if specific patterns are significant. Decreasing the number of intervals results in larger quantization ranges, which decreases the precision of the MUP representation.

It is also worth discussing a hybrid fuzzy logic-based variant of PD developed by Hamilton-Wright (Hamilton-Wright and Stashuk, 2006). In other areas of medicine, fuzzy-logic based inference systems offer an attractive alternative to conventional characterization techniques. The general idea is to represent hard numeric values by linguistic expressions, where a membership function determines the degree to which a variable falls within a particular decision boundary. The benefit of such a system is that it allows quantities to be represented by linguistic qualifiers like somewhat high, high, very high, and expressions such as hot, cold, warm. Hamilton-Wright proposed a fuzzy inference (FI) system that augmented PD with fuzzy logic theory. Quantization error is greatly reduced because values within a bin are assigned memberships based on their position. The 'fuzziness' of membership values allows multiple rules with the same feature value to be considered simultaneously, selecting the rule that yields the best discrimination. Information presented in this way is also more in line with the way humans interpret and reason with data they are presented with, making it an excellent candidate for use in decision support.

3.2.2 Evaluation of MUP Characterization Techniques

In order to be clinically useful, a classifier must meet the requirements laid out in section 2.3.2. As long as the characterization technique is capable of meeting each of these requirements to a satisfactory degree, then transparency can be considered to be the criteria of greatest importance. Conventional classifier methods are simplistic enough that they may provide some level of transparency to clinicians familiar with the concepts of differential diagnosis. However, several drawbacks of these methods, such as unmet assumptions regarding feature data distributions, can limit the accuracy of these methods for MUP categorization. The advanced pattern recognition techniques, such as ANNs and SVMs, are quite robust and offer better performance. However, ANNs are essentially 'black-box' classifiers that do not permit an explanation of their output. Even if such an explanation were possible, it would not be in a form that is easily assimilated by a human decision maker. SVMs also suffer from their complexity and require a deeper appreciation of their theoretical

mechanisms. While providing accuracy comparable to the other methods, the PD and the FI techniques introduced by Pino and Hamilton-Wright offer a completely transparent classification scheme. PD is unique in that it combines information theoretic principles with linguistic interpretation. At the heart of the method lies a robust statistical and probabilistic analysis, but the rules derived from patterns in the data are in a form that can be intuitively understood and easily displayed. In addition, PD and FI are capable of handling continuous and discrete feature values as well as missing data. Further, it is believed that the rule-based nature of PD allows it to generalize its conclusions to new data better than other classification techniques.

3.2.3 Muscle Characterization

Despite the fact that MUPs can be characterized with high accuracy, there is a high degree of variability across MUP characterizations, even for MUPs detected in the same normal muscle, thereby limiting their diagnostic potential. However, a robust muscle characterization can be obtained by aggregating information across a set of MUPs detected from a muscle under test. An overall muscle characterization is achieved by aggregating the characterizations from individual MUPs. Just as in the MUP case, a score is produced for each category under consideration. The muscle is then categorized as being from the class that has the highest conditional probability score.

Confidence

A muscle conditional probability can also be thought of as the confidence in a particular categorization. While confidence is not directly evaluated in this work, it is necessary to understand its basis as it plays a central role in the aggregation of information. A measure of confidence (obtained from the highest muscle characterization score) should reflect the probability that the muscle from which the MUPs were detected is actually of the given category, conditioned on the evidence provided by the set of MUP characterizations. In this way, a muscle characterization measure can be thought of as the confidence in making a particular categorization based on the available evidence. A well calibrated muscle confidence score of 80% for a given category means that, out of all the muscles that are assigned that score, 80% are truly of that category (Pino, 2009). In practice, it is therefore useful to calibrate muscle scores to reflect true conditional probabilities, using a technique such as Monte Carlo simulation (Pino, 2009) or similar method.

Level of Involvement (LOI)

LOI can be thought of in two ways: continuous and discrete. Continuous LOI, as considered in (Pino and Stashuk, 2008), is a single numeric value that is monotonically increasing with respect to the severity of disease. Such a measure can be obtained by analyzing the MUP characterization scores, or the actual feature values relative to exemplary data. For instance, averaging has been shown to produce muscle confidence scores that correlate well with actual LOI, while Bayes' rule tends to saturate to values of 0 or 1 as more evidence is considered. Discrete LOI, on the other hand, is very much like disease detection. Training data is stratified based on various discrete levels of disease severity, such as low, medium, and high. The MUPs of the muscle under test are then characterized with respect to this stratification, and the end result is a characterization vector where the scores represent conditional probabilities for each of the LOI strata. Any of the aggregation measures above could then be used to predict the confidence in this decision, once a particular category is selected. Another principle goal of this work is to determine which of these strategies would be most useful to a clinician.

3.2.4 Aggregation Methods and Decision Making

The most common aggregation methods considered in the literature include the averaging rule (AR) based on the arithmetic mean, and Bayes' Rule (BR), which is consistent with the two main decision making strategies discussed earlier. Pfeiffer and Kunze (Pfeiffer and Kunze, 1995; Pfeiffer, 1999) first introduced the idea of probabilistic characterization by using Bayes' rule for multiple pieces of evidence to aggregate MUPs that were characterized using Fisher's linear discriminant analysis. Pino considered several aggregation metrics, including AR, BR, and the z-transform (Whitlock, 2005). Several other techniques have been developed to aggregate evidence, and are applicable to this framework. For instance, Budescu (Budescu and Yu, 2006) developed and evaluated two separate techniques that attempt to achieve a compromise between Bayes' and averaging. Since the muscle characterization scores generated using BR tend to saturate to 0 or 1 as more evidence (higher number of diseased MUPs) is presented, a rule that is less prone to saturation, while still taking into account the significance of outlier data, is desired.

The rules developed in (Budescu and Yu, 2006) are believed to reflect a decision maker's judgments more accurately on average. In reality, a decision maker's judgment is greatly influenced by the quality and amount of evidence, as discussed previously. However, the strategies employed are

prone to bias and the quality of evidence is not generally known in advance. The two compromising methods will hopefully retain the desired qualities of AR and BR: when the evidence is suspect or weak, the conservative effects will consider all of the evidence equally, and when outliers are present in more than one category, they will be discounted rather than being allowed to saturate the decision towards the wrong category. It is expected that the average accuracy of these methods will fall within the range of accuracy of the AR and BR methods. In fact, the ordering, from most conservative to most extreme is expected to be AR, WMLO, AB, and finally BR. WMLO and AB are expected to have a better balance across classes that are easy or difficult to classify, whereas BR will perform better for well-separated classes and AR will perform better for more challenging class separations.

The aggregation measures considered for decision-making in this work are presented next:

Average Rule (AR)

The simplest aggregation method is to use the arithmetic mean. The AR characterization measure is calculated separately for each category by taking the arithmetic mean of the set of MUP characterization scores per category. Averaging is considered to be conservative because each piece of evidence is inversely weighted by the total amount of evidence. Thus, as more evidence is gathered (in this case MUPs), each MUP characterization is weighted less significantly, reducing the pull of any one particular outlying value. The equation for AR is given as

$$P(y_k|MUP_1, MUP_2, \ldots MUP_n) = \frac{1}{N}\sum_{i=1}^{N} P(MUP_i \mid y_k) \tag{3.1}$$

Where for equations 3.1 through 3.6,

y_k	represents the muscle category (i.e., normal, myopathic, neurogenic),
$P(MUP_i \mid y_k)$	is the probability of MUP_i conditioned on having come from category y_k (i.e. the MUP characterization score assigned to category y_k),
N	is the number of MUPs detected from the muscle under test, and
K	is the number of muscle categories used for classification.

Bayes Rule (BR)

Historically, Bayes has been shown to have better performance in muscle characterization. Typically, when detecting disease, the most significant information is contained in the presence of outliers. This follows from the more conventional techniques such as statistical analysis or even subjective analysis: when more diseased MUPs are present than expected, the clinician often concludes that the muscle is diseased. Thus, BR is additive[2] in that each piece of evidence adds to the total evidence supporting a particular category, and stronger pieces of evidence carry more weight. The equation for BR is given as follows:

$$P(y_k|MUP_1, \dots MUP_n) = \left(\prod_{i=1}^{N} P(MUP_i \mid y_k)\right) \times \left(\sum_{j=1}^{K}\left[\prod_{i=1}^{N} P(MUP_i \mid y_j)\right]\right)^{-1} \tag{3.2}$$

Weighted Mean Log-Odds (WMLO)

The weighted mean log-odds method is essentially an averaging rule. However by considering log-odds prior to averaging, the differences for extreme (low or high) probabilities are accentuated. In this way, evidence that is thought to contain more significant information is given a higher weighting. The equation for WMLO is provided here, however a derivation based on log-odds is provided in Appendix A.

$$P(y_k|MUP_1, MUP_2, \dots MUP_n) = \frac{e^{\beta}}{1 + e^{\beta}} \tag{3.3}$$

Where

$$P(y_k|MUP_1, MUP_2, \dots MUP_n) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\ln(P(MUP_i \mid y_k))}{\ln(1 - P(MUP_i \mid y_k))}\right] \tag{3.4}$$

Adjusted Bayes (AB)

The adjusted Bayes rule is a modified version of BR that includes an adjustment factor. The adjustment factor makes it possible to discount the impact of extreme evidence, essentially pulling in

[2] To visually appreciate the additive nature of BR, the equation must be expressed in terms of *logits*. See Appendix A for an explanation and derivation.

extreme probabilities prior to the application of Bayes rule. A scale-factor of one results in the naïve Bayes rule.

$$P(y_k|MUP_1, MUP_2, \ldots MUP_n) = \left(\prod_{i=1}^{N} P(MUP_i \mid y_k)'\right) \times \left(\sum_{j=1}^{K}\left[\prod_{i=1}^{N} P(MUP_i \mid y_j)'\right]\right)^{-1} \quad (3.5)$$

And

$$P(MUP_i \mid y_k)' = P(MUP_i \mid y_k)^{\lambda} \times \left(\sum_{j=1}^{K}\left[P(MUP_i \mid y_j)\right]^{\lambda}\right)^{-1} \quad (3.6)$$

Where

λ is the risk adjustment parameter that controls the level of discounting of extreme sources of information (Budescu and Yu, 2006). *Note: a value of 0.5 was used for all calculations in this thesis*.

3.2.5 Data Stratification for Confidence and LOI

As a starting point for most QEMG methods, especially those augmented with a CDSS framework, the common approach is to discriminate between broad disease categories such as neurogenic or myopathic, or even the simpler dichotomous states of 'normal' and 'abnormal'. To date, studies suggest that suitable performance can be obtained at this level of decision making, and these methods have been shown superior to conventional methods (Pino, 2009; Pino et al., 2010; Pino et al., 2008). The main goal of this work then, is to expand the granularity or resolution of reported clinical states in an effort to provide more information to the clinician, specifically the level of involvement of a disorder. Obtaining such additional information however, should not compromise accuracy or quality of performance.

The idea of confidence also applies to single MUP characterizations, since these are also conditional probabilities. At this level, the measure of confidence is affected by the amount and quality of the training data. Feature values that fall close to the 'centroid' of a particular set of exemplary data will have higher MUP confidence scores. Such highly confident MUP characterizations can be thought of as outliers or strong pieces of evidence in support of a particular category. Conversely, MUPs with feature values that do not clearly belong to a particular class will be

reflected by lower confidence scores, and their characterization vectors will likely show confusion across categories. An important requirement then, is to have accurate estimates of conditional probabilities.

One approach to attempt to improve conditional probability estimates is to further stratify the training data by discrete levels of involvement, rather than by having a wide range of feature values represented by a single broad category. In training data comprised of differing degrees of disease progression, the overlap of distributions may lead to higher errors, particularly in cases where feature values lie close to the normal/disease class boundaries (as with low LOI). By representing the individual LOI states more concisely, more subtle morphological MUP changes can be detected, leading to better discrimination, and higher levels of confidence in predicting specific LOI categories. However, by increasing the stratification of training data, more data will be needed to represent each class. Also, since there is considerable overlap between the distributions of different LOI categories, performance would be expected to drop regardless of the training set size.

This work introduces a hybrid classification approach that utilizes the advantages of more compact data distributions to increase sensitivity and specificity of disease detection. Off-by-one classification errors are then combined, in order to report the higher level clinical state with greater accuracy and/or confidence. To illustrate the concept, consider a case where a muscle characterization was produced that was divided between myopathic and neurogenic (i.e., had strong support for both disease categories) would still be of clinical utility because it would provide strong evidence for disease versus normality if the evidence was combined. In such a case, confidence for each disease category might be low, but confidence in the presence of disease would still be high. If the clinical goal was to simply detect abnormality, then greater accuracy would be obtained in this case because the classifier had knowledge of the specific disease states. Thus the reference point used in measuring accuracy and confidence play a significant role.

Similarly, estimates of the level of involvement of a disorder would also vary with respect to the chosen reference, and would depend on the class hierarchy or stratification being reported. For instance, if training data were stratified into levels of involvement (e.g., 25%, 50%, 75% or possible, probable, definite), there would be much confusion between these sub-strata due to distribution overlap, but the general detection rate of the underlying disease might be improved by providing examples of specific ranges of abnormality within each class. One goal of this work is to determine a

compromise between category resolution (i.e., the number of clinical LOI states reported) and the accuracy with which each state is detected. A reasonable trade-off of category resolution and characterization accuracy would facilitate the reporting of clinically useful levels of involvements.

Chapter 4
Data Used for Evaluation and Clinical Validation

Simulated EMG data was used to develop and test the PMC framework. The results were then validated by repeating a subset of the tests on real world data. This section first describes the data sets used, and then discusses descriptive statistics and presents feature distributions that may give an indication as to the ease or difficulty with which the data can be classified.

4.1 Data used for Evaluation

To form a basis for discussion in evaluating the classification and aggregation techniques at different degrees of disease progression (i.e., levels of involvement), EMG data was simulated according to an electrophysiological model (Stashuk, 1993). The simulator was used to create pools of data that are representative of the main categories of neuromuscular disease (i.e., healthy, myopathic and neurogenic). To simulate a neuropathy, motor units are randomly atrophied, while orphaned muscle fibers are randomly re-innervated by nearby surviving motor neurons that are within a defined range. To simulate myopathy, a percentage of healthy fibers are 'infected' in a manner where some fibers are atrophied (in diameter) to a small degree while others are hypertrophied to an even smaller degree. This process is iterated to successively infect fibers to a larger degree, while at the same time infecting new fibers, until a prescribed level of involvement is reached. A fiber is considered non-functioning (i.e., dead) when its diameter is below a critical threshold. Muscle activation is regulated by specifying the percentage of maximal voluntary contraction (MVC) force generated in a healthy muscle. If a disease process is present, the number of motor units recruited and their firing patterns are either increased or decreased, to achieve the same level of force.

The simulated muscle signals were modeled to have been detected using a concentric needle electrode that sampled the muscle at various intramuscular locations within a 2mm square centered about the muscle model's center. All contractions were acquired at 10% MVC and were decomposed into constituent MUP trains using DQEMG (Stashuk, 2001) software.

When stratified by disease category, the myopathic and neurogenic MUP pools were comprised of about 1000 MUPs each. The control MUP pool had a total of 1500 MUPs. The stratification of MUPs into disease categories (CTRL, MYO, and NEUR) represented the lowest resolution of classification available. The myopathic and neurogenic MUPs were simulated to come from muscles with 25%,

50% and 75% muscle-fiber/motor-unit loss, in approximately equal proportion (approximately 300 MUPs in each LOI pool). Each pool of diseased MUPs was further stratified by LOI (labeled MYO-25, MYO-50, MYO-75, NEUR-25, NEUR-50, and NEUR-75). This stratification represented the highest resolution of classification available, with the CTRL group included as the seventh class.

4.2 Data Used for Validation

Data was obtained from a clinical source in order to validate the findings that were evaluated using the abovementioned simulated data. However, the dataset represents only a single disease state (myopathic), and is stratified into two levels of involvement rather than three. This poses several limitations in applying the methods developed in this work.

4.2.1 Muscular Dystrophy Data (MDX)

Disease and control data collected at the London Health Sciences Centre (London, Ontario) comprised of several types of muscular dystrophy at various levels of severity, as well as a set of control subjects. In total, there were four patients with facioscapulohumeral dystrophy (FSHD), six patients with limb-girdle muscular dystrophy (LGMD), and five patients with Becker's muscular dystrophy. The control group consisted of seven subjects.

The study was comprised of fifteen individuals, with a previously determined diagnosis of muscular dystrophy, and a group of seven healthy subjects with no evidence of neuromuscular or musculoskeletal disorders. Participants diagnosed with muscular dystrophy were recruited through the Neuromuscular Clinic at University Hospital, London Health Sciences Centre, London, Ontario. Diagnosis was confirmed by clinical assessment and genetic testing. Control subjects consisted of recreationally active individuals recruited from the university environment.

Among other aspects of the protocol, each patient's maximum voluntary contraction (MVC) of the quadriceps muscle was recorded via a torque gauge and oscilloscope. Needle EMG was recorded from the biceps, vastus lateralis, and tibialis anterior muscles of each patient. Patients from the disease group were arbitrarily divided into equal groups of low and high levels of severity based on their quadriceps MVC measures.

The lowest resolution of classification was obtained when MUPs were stratified by disease, resulting in two categories or clinical states (CTRL, MYO). The MYO pool was further stratified by

LOI (labeled MYO-L, MYO-H). This stratification represented the highest resolution of classification available, with the CTRL group included as the third class.

4.3 Statistical Analysis of the MDX Data Set

A preliminary analysis of each of the data sets was conducted in order to gain a qualitative appreciation of the underlying distributions of data. Selected feature distributions were plotted using a Parzen window estimation technique and Gaussian kernel (Duda et al., 2000)[3]. The density estimate was evaluated at 100 equally spaced points covering the range of the feature data. Feature densities are plotted in Figure 4.1 through Figure 4.9 for the control group and for each LOI group on a single plot per disease category.

4.3.1 Analysis of Feature Distributions

Distribution plots of selected features are presented for discussion. As discussed in the background section, certain MUP features, such as phases and turns, are non-specific and increase with neuromuscular disease. However, other features, relating to size and shape, have specific trends that depend on the underlying disease process. A plot of area density for each level of severity illustrates these trends. In a neurogenic disease process, area increases with LOI and decreases with LOI in a myopathic disease process, as illustrated in Figure 4.1 and Figure 4.2. Similar trends are observed for thickness and duration in the neurogenic case. The trends seen in myopathic data are validated using the MDX data set (refer to Figure 4.3). For the most part, the number of turns and phases either increase, or stays relatively the same.

Two notable exceptions are seen with turns, thickness and duration in the myopathic case, particularly with simulated data. First, the distribution of turns is more concentrated around a value of three for normal MUPs, and increases with disease (see Figure 4.4 and Figure 4.5). However, it is interesting to note that there are fewer MYO-75 muscles with very high turns, suggesting that as LOI increases, the likelihood of finding MUPs with higher numbers of turns actually diminishes, and tends toward the normal category. This is contrary to the expectation that complexity should grow with disease severity. Thickness also shows an odd trend, and seems to shift to the right (increase) for lower levels of myopathy, but as the level of disease increases, the range of thickness values widens

[3] The window parameter, u, was calculated using the expression, $u = s*(4/(3*N))^{(1/5)}$, where s is the sample standard deviation and N is the total number of samples in the feature data.

due to an increase in the likelihood of abnormally thin MUPs (see Figure 4.6 and Figure 4.7). Although the initial increase is not expected, a reduction of thickness with myopathy is typical.

Duration was also found to have an uncharacteristic initial tendency to increase, and then decrease with higher LOI as the likelihood of shorter MUPs increases, but not to the degree that would be expected (Figure 4.8). Similar trends were not observed in the MDX data (Figure 4.9), except in the case of turns where there is a higher likelihood of lower turn values for the severe group.

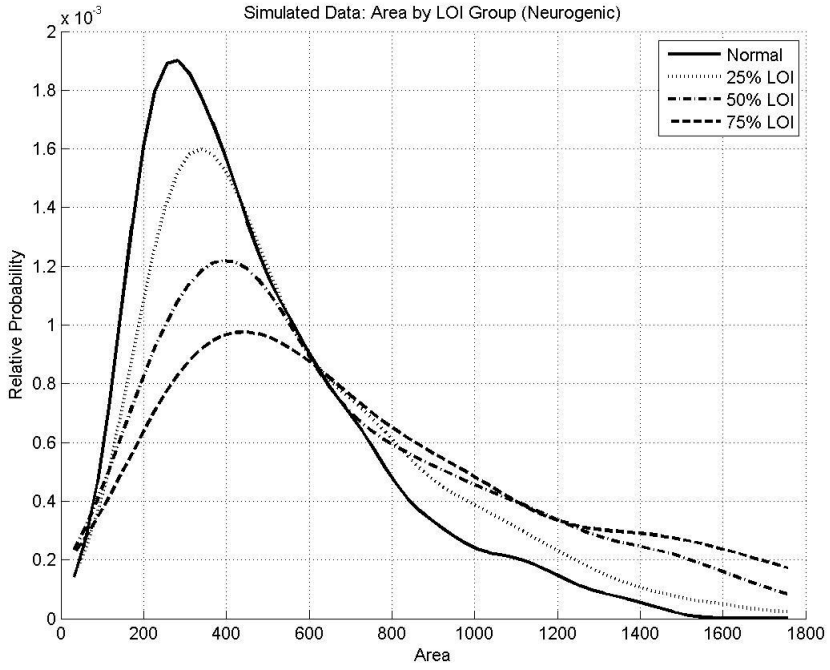

Figure 4.1: Estimated distribution of area - simulated neurogenic data

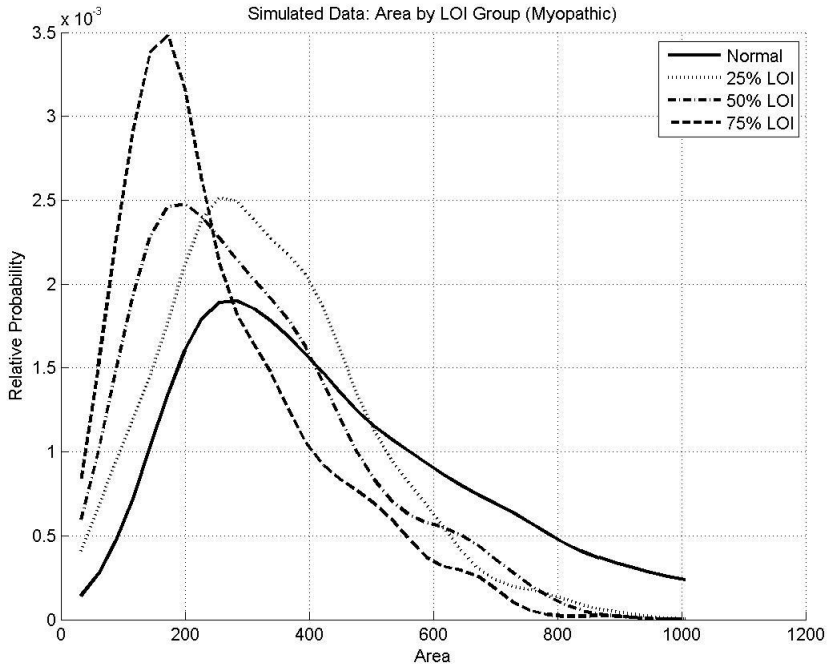

Figure 4.2: Estimated distribution of area - simulated myopathic data

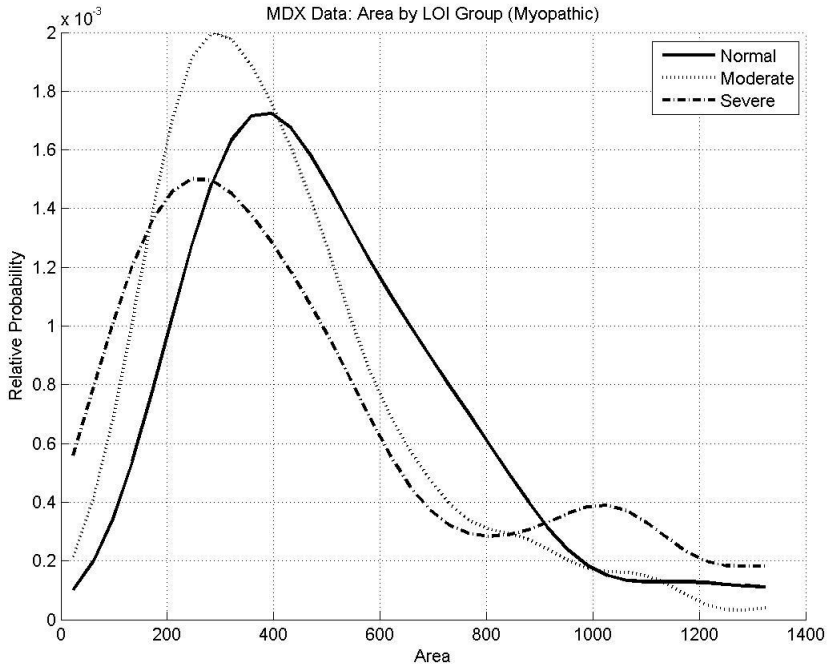

Figure 4.3: Estimated distribution of area - MDX myopathic data

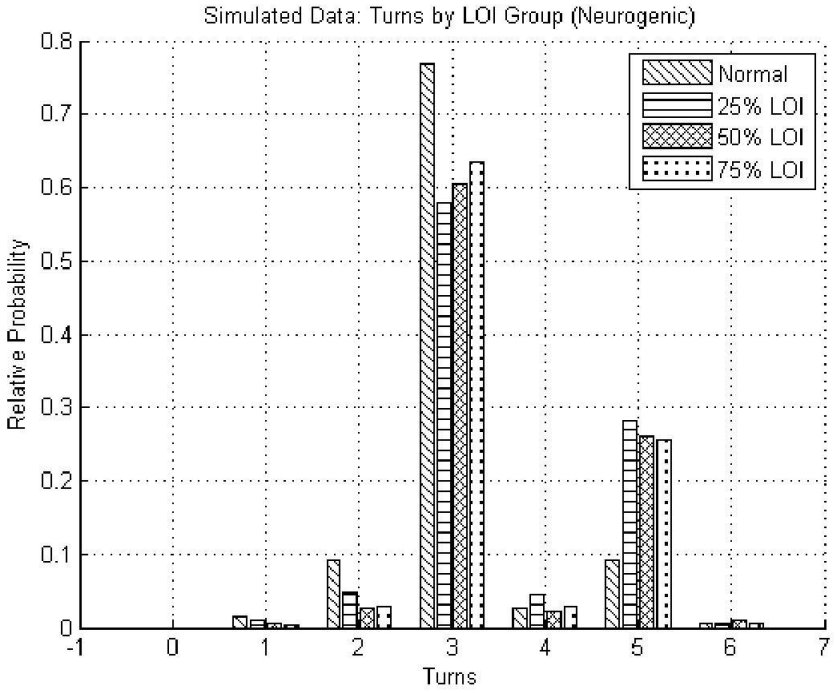

Figure 4.4: Estimated distribution of turns - simulated neurogenic data

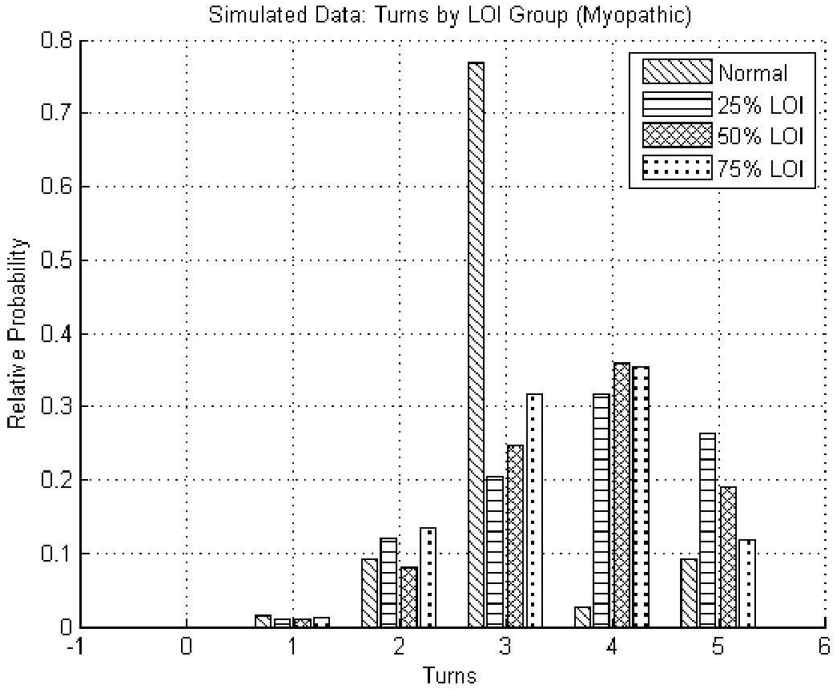

Figure 4.5: Estimated distribution of turns - simulated myopathic data

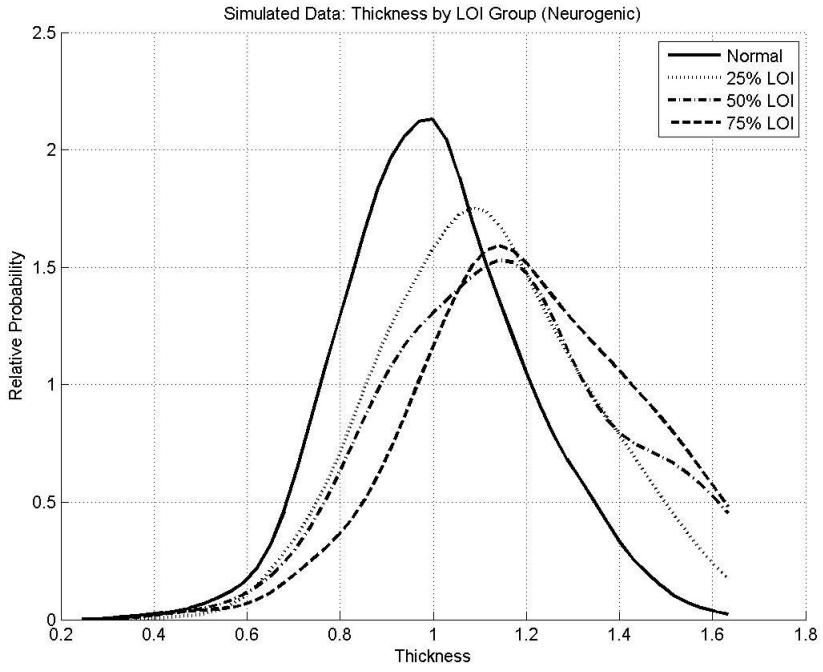

Figure 4.6: Estimated distribution of thickness - simulated neurogenic data

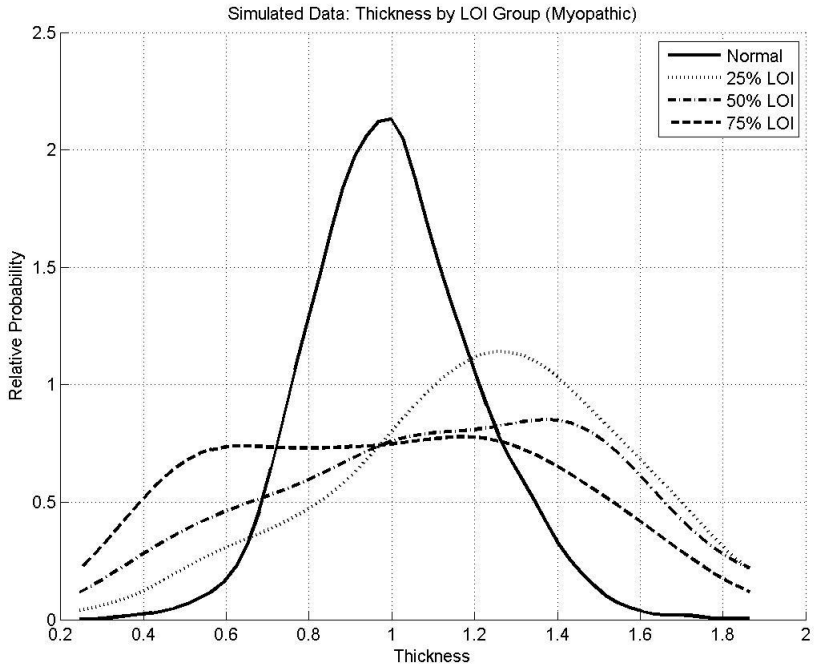

Figure 4.7: Estimated distribution of thickness - simulated myopathic data

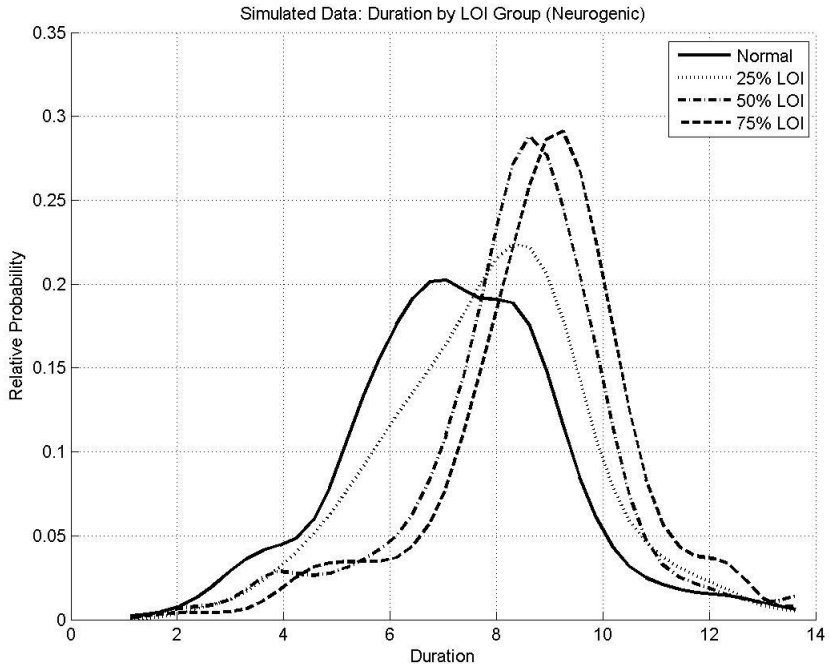

Figure 4.8: Estimated distribution of duration - simulated neurogenic data

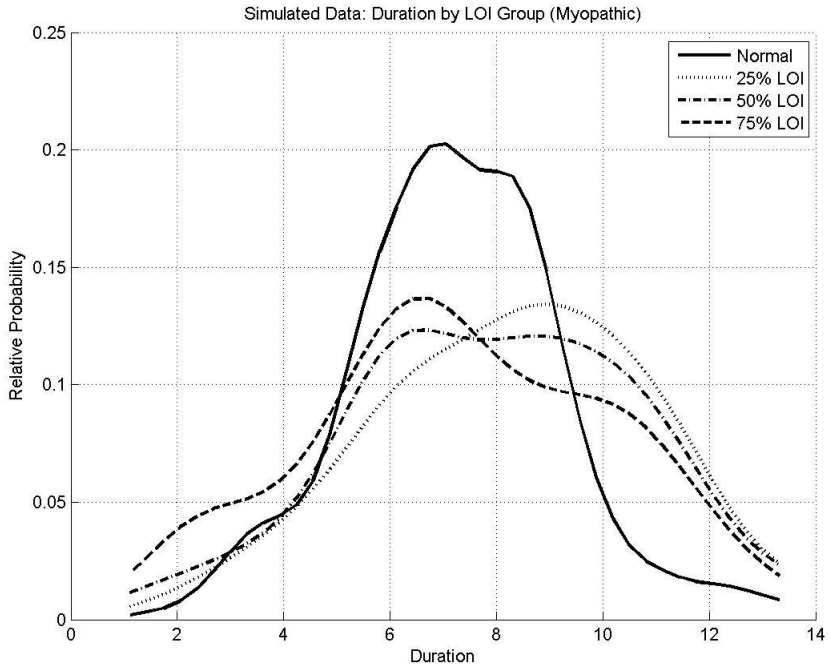

Figure 4.9: Estimated distribution of duration - simulated myopathic data

4.3.2 Statistical Analysis of Clinical Data

Previous work (Derry, 2010) applied statistical hypothesis testing to the MDX data to determine which, if any, of the average MUP features were significantly different across LOI and control groups. The findings are summarized in this section.

4.3.2.1 MDX Data

Statistical analysis performed by the researchers involved in this study showed significant differences between diseased muscle groups and control subjects for MUP amplitude, AAR and turns in one of more of the three muscle groups (e.g., VL: $p < 0.0001$, TA: $p < 0.05$, BB: $p < 0.0001$), but no significant differences for the average number of phases (Derry, 2010).

In the vastus lateralis, MUP duration, area, and AAR were significantly different (reduced) in the severe group compared with the mild group ($p < 0.0001$), however no significant differences were observed between groups for amplitude, number of phases or turns. For the tibialis anterior, only the number of turns was found to be significantly different ($p < 0.01$), and for the biceps brachii both turns and phases were significantly different ($p < 0.0001$). No significant differences were noted between the mild MD and control groups (Derry, 2010).

4.4 Discussion

For most of the features discussed, the expected trends are observed in both the simulated data and clinical data sets. However, the notable exceptions occurring in simulated myopathic data for turns, duration and thickness can suggest several things. In particular, since these trends were not confirmed in the MD data set for duration and thickness, there are two possibilities. First, the simulator model of myopathy may be inaccurate for low levels of myopathy, in that the ratio of fiber death to fiber infection may be too low. Thus for low levels of involvement, a large number of inactive fibers would add a significant delay in conduction of the MFP, resulting in abnormally longer durations. A second possibility is that the MD data was collected from patients who had higher levels of disease severity. As such, any early restructuring that could explain these distribution patterns was no longer present, and fiber atrophy was already the predominant effect. In either case, some degree of discrepancy between the model for myopathy and the actual disease process is likely present.

The distribution of turns suggests that there is simply a higher variance in turns as LOI increases. While this is not useful on its own, a classifier such as PD might be able to detect useful patterns such as low area and increased or decreased turns.

Distributions for the remaining muscles of the MD data set were not evaluated. Since the vastus lateralis (VL) muscle was used to measure MVC, and it is that MVC measure that was used to stratify the data, it is expected that the best separation among classes would be seen in the VL group.

The distributions presented and discussed show that there clearly are observable changes between control and disease groups, and further between various LOI groups. However, conventional statistical analysis suggests that the separation between low LOI groups and control groups is not significant. Further, when discussing single features, the limitations in predicting the likelihood of a particular category based on MUP feature values are quite obvious. It is for these reasons that the probabilistic techniques presented in this thesis have a significant advantage. The first advantage stems from the fact that these methods employ multivariate statistics, allowing simultaneous consideration of several features and increasing the richness of information available. The second possible advantage stems from the fact that non-linear trends in feature distributions can lead to unique patterns in the data. A classifier such as PD would be able to detect these patterns, and with sufficient training samples, should outperform conventional linear classifiers such as LDA. The wide ranges of values present in these data also suggest that standard disease-based stratification would have some limitations, particularly with lower LOI. LOI-based stratification presents further problems because of the amount of overlap between LOI group feature distributions. Thus the hierarchical techniques that will be considered in this work will attempt to address these limitations. By classifying data with the highest level of granularity or resolution, and then reducing this resolution to the desired level, it is possible to capture the subtle changes in MUP morphology that are essential for discrimination. This is of most importance for low LOI, because many of the MUP features have significant overlap with the control group. If these feature values in this range are not properly accounted for, they will be misclassified more often.

Chapter 5
Methods

The previous chapter described current implementations of the PMC framework and identified several areas for improvement. Namely, current methods perform reliably when detecting disease in muscles with high disease LOI but tend to perform poorly at detecting disease in muscles with low LOI. A hierarchical classifier is developed and compared with existing methods with the objective of improving the balance of accuracy across LOI groups. To further the utility of quantitative methods, a second objective is to provide a clinically useful measure of LOI in the form of clinical LOI states. A variation of the hierarchical method is considered to determine the optimal resolution of LOI states that can be reported with reasonably high accuracy. Continuous measures of LOI are also considered, since previous techniques have demonstrated that although muscles correlate well with LOI, the wide range of scores within each LOI group makes it difficult to measure LOI reliably. Techniques are developed to attempt to reduce this variability. Finally, to understand the impact of different decision-making strategies (conservative versus extreme), the behavior of the aggregation measures introduced previously is studied as the amount of evidence is varied. In particular, the number of MUPs considered in characterizing the muscle is increased at specified intervals. The purpose of the latter analysis is to deterministically describe the conditions that affect the performance of each aggregation measure, so that appropriate measures can be used to optimize global performance.

5.1 Clinical States for Diagnosis

For the simulated data used for evaluation purposes, there were three clinical *states* on which to base a diagnosis, namely normal (CTRL), myopathic (MYO), and neurogenic (NEUR). These states represented the highest level of stratification, or the coarsest stratification resolution that was possible for classification. Other studies have considered an even more general stratification using only two classes (e.g., normal vs. abnormal) (Katsis et al., 2007; Stålberg et al., 1994; Stewart et al., 1989). It is possible that such higher level stratification would be desirable from a workflow perspective, but would not necessarily lead to improved performance. For each disease category, the data was further stratified by simulated level of involvement, resulting in low (L), medium (M) and high (H) LOI *groups*. This represented the lowest level of stratification and subsequently finest level of resolution

available for classification. The CTRL group was not sub-divided in such a manner, thus in total there were seven high-resolution classes.

5.1.1 Data Stratification and Terminology

To provide clarity for the reader, efforts have been made to use consistent terminology wherever possible. Perhaps the most confusion lies in explaining the ideas of data stratification. Strictly speaking, when discussing how data is stratified for the purposes of classification, individual pools of training data are assigned to a particular *category* using a *class* label. Thus, these terms may be used interchangeably to refer to the class-structure of a training set. The term *category* is also specifically used to refer to particular strata of disease or LOI group, or to describe a clinical state presented to a decision maker. Following this logic, a *characterization* is the set of scores assigned to each *category*, and can be applied to a muscle or an individual MUPT. Although the muscles used for testing are also stratified by LOI category, they are referred to as *groups* (i.e., control group, disease group, etc). The individual groups may also be assigned to a category label, and in some cases (as in the hierarchical classifier scheme) two or more groups may be combined and assigned to the same label (i.e., the 25-, 50- and 75% myopathic *groups* are assigned to the myopathic *category*). Even if the groups are labeled in this way, accuracy measures (sensitivity) are presented on a per-group basis, to allow for better evaluation of the consistency and robustness of each method. For instance, a method that performs perfectly for severely diseased patients but fails in detecting pathology in mildly diseased patients is not desirable. A method that performs similarly in each muscle group would presumably provide a better balance of performance and consistency, even if the average accuracy was the same. (Note: in this sense 'balance' is quantified by the sensitivity-specificity deviation (SSD) measure introduced later in this chapter.)

Most models assume that the categories used for classification are the same as those presented to the decision maker (e.g., the conventional disease states), however the hierarchical model presented in the next section precludes this. Thus, discussions relating to the categories or class structure used to stratify training data will refer to the *level of classification*, whereas discussions relating to the clinical categories presented to a decision maker will refer to the *level of inference* (refer to Figure 5.1). The strata used at these two levels may or may not be the same. In the case of disease detection, the clinical states corresponding to the lowest stratification resolution (i.e., disease categories) were always reported at the inference level. However, at the classification level, different degrees of

resolution were considered. In fact, the main classification strategy developed can be viewed as a high resolution classification problem that is then reduced to the resolution required to represent the clinical states being studied. In the case of LOI detection, the appropriate resolution of clinical states to report at the inference level was unknown. This was one of the parameters that were varied, and the goal was to find a balance between accuracy and LOI resolution. The specific strategies and permutations of stratification groups for each of these cases will be described in more detail in the next sections.

5.2 Characterization Strategies

As a starting point, training and testing data were evaluated using both high-level disease stratification and low-level LOI stratification. Accuracy measures were based on the aforementioned stratification schemes, resulting in square confusion matrices. MUP feature vectors were classified using LDA and PD, and were aggregated using the four aggregation measures described: Arithmetic Mean (AM), Bayes Rule (BR), Adjusted Bayes Rule (AB), and WMLO.

5.2.1 Accuracy of Disease Categorization

Previous preliminary testing showed that errors made during low-level stratification were typically made within disease groups. As a result, a hierarchical classification scheme was developed. At the lowest level, each MUP is classified using a high class resolution, treating each LOI sub-strata as a separate class. The size of (number of conditional probabilities in) the resulting characterization vector is then equal to the number of LOI groups multiplied by the number of disease groups, in addition to the CTRL group. MUP characterizations of this form are then aggregated using one of the techniques mentioned above, producing a muscle characterization vector of the same size. The resolution of this vector is then reduced by combining the scores assigned to each of the LOI groups of a specific disease state. The score assigned to the CTRL category remains unchanged. In the current implementation, the scores corresponding to each of the LOI groups are combined by summation, and thus remain normalized, although they no longer reflect true conditional probabilities. The top portion of Figure 5.1 illustrates the hierarchical classification scheme.

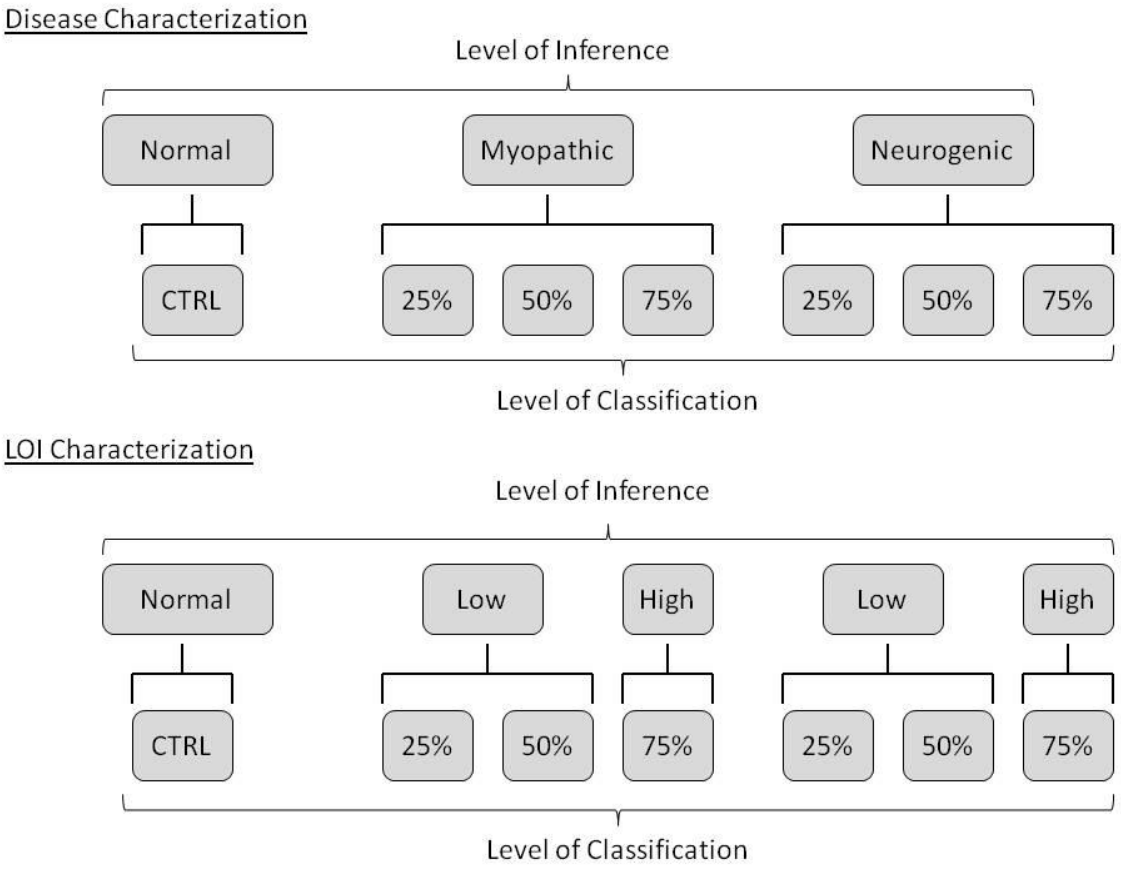

Figure 5.1: Hierarchical classification scheme applied to disease and LOI characterization

5.2.2 Accuracy of LOI Categorization

LOI can either be classified into discrete categories such as low and high, or it can be reported as a continuous value that relates to the confidence in a characterization. Since LOI is a relatively new concept in clinical QEMG, there is no gold standard or preferred way to present the information. Thus, the purpose of these methods is not only to improve accuracy for a particular method, but to compare the various strategies in terms of their accuracy and resolution, so that recommendations can be made. The first part of this section refers to identification of discrete LOI classes, which is analogous to the disease detection process described above. The second part attempts to evaluate continuous measures of LOI. There is some degree of overlap between the two strategies, since the discrete methods are also modified to produce a continuous score that relates to LOI.

Discrete LOI Categories

One of the issues in using the high-resolution stratification above is the relatively large number of classes (i.e., 7 vs. 3). While the technique may improve upon the accuracy of disease detection, an increase in the number of classes requires a subsequent increase in the amount of total training data, as a sufficient number of samples is needed to represent each category. PD was shown to be particularly affected by training set size, and when the training pool is of insufficient size, PD is less accurate at estimating conditional probabilities. The underlying premise here is that accurate conditional probabilities are more important in determining LOI, especially in the case of continuous measures, but also in the case of discrete LOI classification.

Thus, in evaluating LOI methods, a two-stage classifier was introduced. The first stage used the hierarchical scheme described in the previous section to determine the disease category. The second stage (LOI stage) then focused on predicting the LOI category corresponding to the predicted disease category only. The second stage only required training data for the disease category that was predicted in the first stage, thus reducing the number of classes to be considered. The addition of the control group to the training pool used in the second stage was also considered.

For discrete LOI classification, several variations of the hierarchical scheme were applied in the LOI stage. The first set of techniques focused solely on relabeling and reassigning classes, in a manner similar to the hierarchical scheme used for disease detection. Preliminary testing also showed that many of the errors in predicting LOI categories were made by incorrectly assigning muscles from the medium (M) LOI category as either low (L) or high (H). Thus a re-assignment strategy was applied that would re-label muscles originally classified as M based on the highest characterization score assigned to each of the L and H categories. In other words, if the M category was characterized with the highest score, then the second highest score was considered in order to assign a category.

Consequently, the LOI category resolution provided at the inference level was reduced from three categories (L, M, and H) to two categories (L, H). In order to evaluate performance and to construct a confusion matrix, it was thus necessary to merge the true labels of the test muscles so that there were only two categories, namely L and H. Two ways to merge the labels were considered, the first combined the muscles of the M group with the L group, while the second combined the muscles of the M group with the H group.

The second set of techniques focused on relabeling categories at the level of classification. For these techniques, the class resolution of the actual training data was reduced. Just as in the first case, the M group was either relabeled to be part of the L group, or relabeled to become part of the H group. By relabeling at the classification level, MUP and muscle characterization vectors already represented the actual clinical LOI states, and it was not necessary to reassign predicted categories as before. The true labels of the test muscles were again merged as described above.

The main differences between the first and second sets of techniques is that in the former, the resolution of categories was reduced, while the classifiers still relied on a fine-resolution class structure and in the latter, relabeling at the classification level resulted in training the classifier with a coarser class-structure resolution. These differences, along with a visual depiction of the hierarchical stratifications, are illustrated in Figure 5.2.

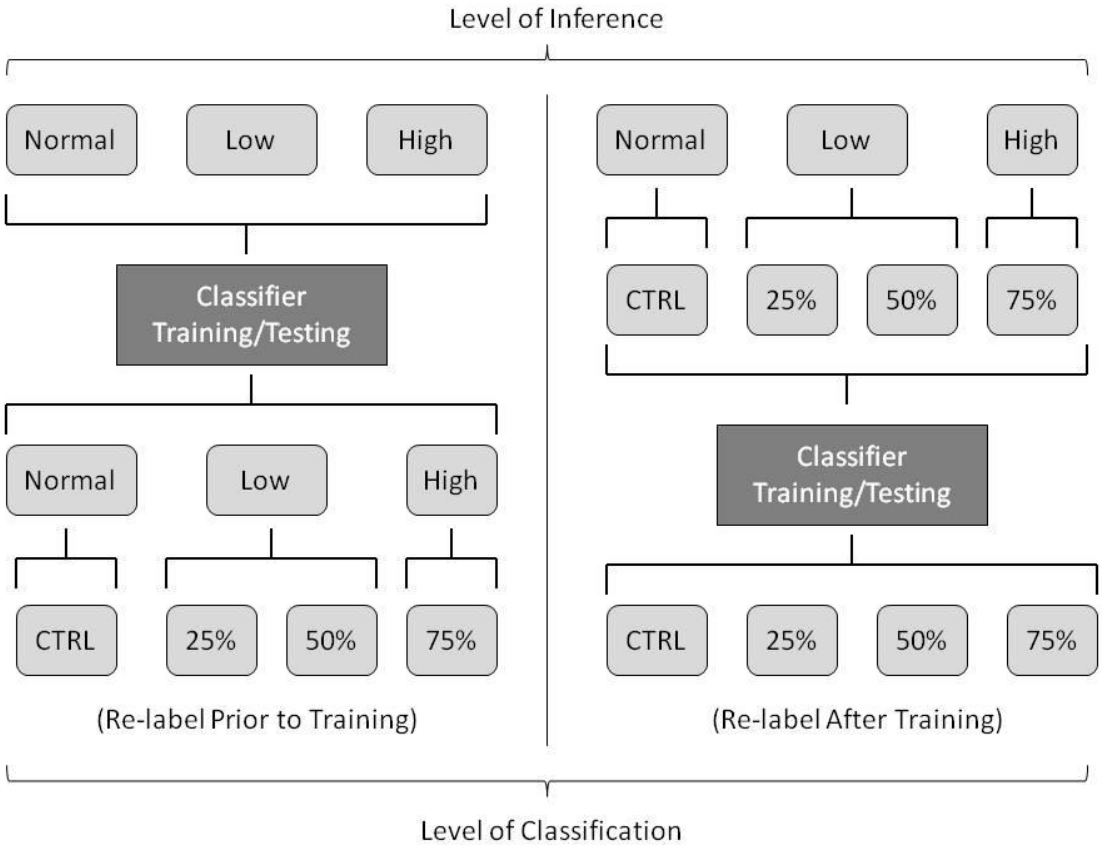

Figure 5.2: Differences between hierarchical stratification schemes for LOI characterization

Continuous LOI

In addition to evaluating discrete LOI characterization performance, continuous measures relating to LOI were considered. Such evaluation was first introduced by Pino (Pino and Stashuk, 2008), who demonstrated that characterization scores of the winning category were highly correlated with LOI. However in that previous work, only the AM aggregation technique was evaluated. In this work, scores produced by AM, BR, AB and WMLO were considered. In the case of disease stratification, winning characterization scores were correlated with LOI for the purposes of comparison against the findings discussed in (Pino and Stashuk, 2008). However, these scores were shown to have considerable overlap and were not found to be clinically useful. As such, the stratification and classification schemes considered in this work were modified to produce a continuous scalar value representing LOI.

The first method considered involved the two-stage classifier scheme introduced above. This technique was developed purely as a continuous LOI measure, and did not attempt to classify and assign a label based on LOI category. In the LOI stage of classification, a classifier was trained using the MUP data corresponding to the predicted disease category, in addition to the CTRL data. The disease data was not stratified any further. Thus the objective was to use the conditional probability assigned by the binary classification problem as a measure of LOI. If the predicted category was normal, then the second stage was bypassed and a score of zero was assigned for LOI. However, in the case of a predicted disease, the conditional probability score assigned to that category by the second stage LOI classifier was used, regardless of whether it was the winning score.

For the discrete LOI classifier variants, it was necessary to apply an interpolation scheme or heuristic to obtain a continuous score that was a combination of the characterization scores produced by the aggregation process. Two types of scores produced by the hierarchical classifier were considered. The first method obtained LOI scores by summing the conditional probability measures assigned to each LOI category at the level of classification, while the second method used an interpolation scheme, also based on these conditional probability measures. The former was only applied for completeness so that the hierarchical classifier score could be compared with the standard disease-stratified and LOI-stratified techniques.

The interpolation method was used in order to obtain a more intuitive continuous score, and can be referred to as deterministic mapping, rather than a conditional probability or measure of confidence.

If the number of conditional probability measures in the characterization vector was odd, the middle category was used as the reference, which was assigned a value of zero. Scores to the left of this value (i.e., scores assigned to categories representing lower LOI groups or the normal category) were subtracted from the reference, while scores to the right of this value (i.e., scores assigned to categories representing higher LOI groups) were added to the reference. The obtained value was then normalized by the number of classes to obtain a continuous score in the range of [0 1]. If the number of conditional probability measures in the characterization vector was even, the scores were split evenly and the midway point between the two middle categories was chosen as the reference. The reference value was set to zero as before. Scores assigned to categories representing lower LOI groups were subtracted from the reference value and scores assigned to categories representing higher LOI groups were added to the reference value. The value obtained was again normalized by the number of LOI groups considered. Equation 5.1 shows the interpolation scheme used in each of these cases.

$$LOI = \sum_{w=-W/2}^{W/2} w \cdot P(y_w|MUPs) \tag{5.1}$$

Where

$$W = \begin{cases} number\ of\ categories & if\ even\ (w \neq 0), \\ number\ of\ categories - 1 & if\ odd, \end{cases}$$

$P(y_w \mid MUPs)$ is the characterization score assigned to the w^{th} category, when scores are sorted by increasing LOI.

When high-resolution stratification was used at the classification level, the same interpolation scheme was applied regardless of how the characterization was reported at the inference level. However, when muscle groups were relabeled at the classifier level, the size of the characterization vector was reduced, and so the interpolated continuous values were also dependent on class stratification. In both cases, the score was evaluated with and without the CTRL group data included in the training set.

5.2.3 Evaluation of Aggregation Measures

For all of the above techniques, each of the aggregation measures was applied so that their relative performance could be evaluated. However, to study the behavior of the various aggregation measures it was necessary to vary the amount of evidence used in arriving at a decision. Thus the number of MUPs used for aggregation was varied, in increments of five, starting with five MUPs and ending with 25 MUPs. The accuracy and SSD of each method was evaluated for each set of muscles comprised of the specified number of MUPs.

5.2.4 Validation of Methods Using Clinical Data

The above techniques were applied to simulated data for evaluation purposes. The findings were then validated against the clinical MDX data described previously. LOI detection was only evaluated for certain methods because the hierarchical model could not be used. For the MDX data, since there were only two LOI categories, it was not possible to combine them and still provide an LOI characterization. Continuous LOI correlation was still considered.

5.3 Evaluation of Performance

The following section describes the measures used to evaluate the performance of methods used in this thesis.

5.3.1 Accuracy and Balance between Sensitivity and Specificity

Measures of performance that are more robust to unequal numbers of test instances across classes exist (Pino, 2009). For instance, Pino used sensitivity, specificity and sensitivity-specificity deviation (SSD) to evaluate classifier performance. SSD is used to determine how well a classifier maximizes (or balances) both specificity and sensitivity. Traditional measures of accuracy are biased towards the category that has the largest number of test muscles to be characterized (Pino, 2009). Thus, measures of accuracy become skewed when the proportion of disease muscles to controls is unequal.

Sensitivity was defined as the total number of muscles characterized as having a particular disease divided by the total number of ‘true’ muscles having that disease. Specificity was defined as the total number of muscles characterized as normal divided by the total number of ‘true’ normal (CTRL) muscles.

In this work, the traditional measure of accuracy is used within each disease and LOI strata. When considered in this way, accuracy is equivalent to sensitivity when the muscle is from a disease pool, and is equivalent to specificity when the muscle is from the control pool. However, since there are multiple disease-LOI strata, a mean accuracy is computed across all such measures, which is also used in the calculation for SSD. Sensitivity, specificity and SSD are reported when there is only one disease category, and a 3-class SSD is used in the a 3-class case.

Total accuracy was defined as the average of sensitivity and specificity. If there was more than one disease category, then sensitivity measures from both disease categories were included in this average.

In the two-class case, SSD was defined as:

$$SSD = \sqrt{\frac{(A - sensitivity)^2 + (A - specificity)^2}{2}} \tag{5.2}$$

For data sets with three categories, the term sensitivity-specificity deviation (SSD) was defined as:

$$SSD = \sqrt{\frac{\left((A - A_{norm})^2 + \left(A - A_{myo}\right)^2 + (A - A_{neur})^2\right)}{3}} \tag{5.3}$$

Where

A_{norm} is equivalent to the specificity term in equation 5.2 (i.e., normal class accuracy),

A_{myo} is the mean sensitivity across each LOI group within the myopathic category,

A_{neur} is the mean sensitivity across each LOI group within the neurogenic category,

A is the mean accuracy of A_{norm}, A_{myo}, and A_{neur}.

5.3.2 Correlation of Muscle Scores with LOI

Muscle characterization scores were assessed for correlation with actual LOI using Spearman Ranking (Duda et al., 2000), and were calculated for all characterized muscles using the score assigned to the true class, regardless of whether that class was correctly categorized with respect to disease. However, the muscles of each disease group were considered separately, and muscles from the CTRL (normal) group were not included in the correlation measure. The effect of including muscles from the CTRL group in the training set used for classification was assessed. By including

muscle scores from muscles that were incorrectly categorized, correlation values may be skewed, but it seems logical to allow these mistakes since in practice it is a plausible scenario and correlation assessments are not concerned with categorization accuracy. In addition to recording correlation values, plots of muscle score vs. actual LOI are generated showing individual score values.

Chapter 6
Results

The results of all tests are presented in this chapter. First, the sampling protocol is described as it pertains to the datasets used for evaluation and validation of the methods. Next, results pertaining to the improvement of disease categorization are provided, to determine if the hierarchical approach is able to provide better accuracy and balance across LOI groups. The analysis of methods used to categorize LOI into discrete classes is then presented. Here, different class hierarchies and groupings are considered in order to determine the ideal balance between the resolution of LOI categories and the accuracy with which those categories can be detected. Continuous measures of LOI produced by the aggregation measures described previously are then considered using the various stratification schemes, to determine if a particular method accurately represents LOI. Finally, the results describing the behaviors of the various aggregators are presented so that the suitability and necessary conditions of each aggregation method can be evaluated.

6.1 Sampling and Presentation

As discussed previously, the data used for evaluation purposes was comprised of a control state and two disease states, and each of the disease states was further stratified into three LOI groups, totaling seven groups. For validation purposes, clinical muscular dystrophy data was used in which there was only one disease state and two LOI groups, totaling three groups.

6.1.1 Sampling of Data for Evaluation

In total, 700 virtual muscles were sampled from the MUP pools for training and testing purposes. Each virtual muscle was comprised of a specific number of MUPs and was assigned a category label corresponding to the pool from which its MUPs were drawn from. For each of the LOI MUP pools, and for the CTRL MUP pool, 100 muscles were created using a *k*-fold cross-validation scheme (k = 10). For every k^{th} fold, the MUP pool was roughly split in half, with 150 MUPs randomly selected (without replacement) for training. From the remaining MUPs, *m* virtual muscles (m = 10) were created for testing by randomly selecting *n* MUPs per muscle without replacement, where *n* was the number of MUPs used for the particular study (in most cases n = 10). The set of *n* MUPs selected for the m^{th} muscle was replaced before drawing subsequent muscles. This process was applied to each of the MUP pools, creating 10 muscles and a set of training data for each high-resolution stratification

group. This entire selection process was repeated for a total of *k* folds (k = 10), thus producing 100 muscles for each stratification group, and 10 unique sets of training data (i.e., one per fold). To study variation across training sets, the muscles from each cross validation group were combined, but the training data corresponding to each of the *k* folds were kept separate, so that the muscles drawn from each of the *k* cross-validation groups were trained with unique training data.

The same set of virtual muscles was classified using LDA and PD, and their individual MUP characterizations were combined using the AM, BR, AB and WMLO aggregation measures. For each of the desired study objectives (i.e., Disease detection, LOI detection) the same set of virtual muscles was used across the various stratification schemes as well. In this way, the variability across trials was minimized. An exception was made to this method rule when studying the effects of the amount of evidence (i.e., number of MUPs) on aggregation techniques because of the variations in training set size produced by these protocols.

6.1.2 Sampling of Data for Validation

To validate the findings on real world clinical data, the methods herein were applied to the MDX data set. This data set has a relatively low number of muscles, especially when stratified by LOI. As such, it would be difficult to assess characterization performance based on the actual muscles because accuracy measures would be very granular and variability would be high. The 10-fold cross validation scheme described above was thus applied to these data, creating unique training pools and virtual muscles for testing. For each of the LOI stratification categories, as well as the control group, 100 virtual muscles were created, for a total of 300 muscles.

6.2 Accuracy of Disease Categorization

The performance results of the hierarchical classification scheme are compared with the standard classification methods. Both total accuracy and performance across LOI groups are considered, in order to evaluate accuracy and the ability to balance sensitivity and specificity.

6.2.1 Evaluation Using Simulated Data

In general, total accuracy for the simulated data set was exceptionally high. Using disease stratification, high accuracy was achieved when muscles were comprised of either 10 or 20 MUPs. With 10 MUPs, LDA and PD were comparable with accuracies between 94-96%. LDA had a significantly lower SSD that was more than half the value of the SSD for PD. These results are

illustrated in Table **6.1** for the case when 10 MUPs were used. Although not shown, at 20 MUPs, PD accuracy increased to 97% with a decreased SSD (0.04-0.08) and LDA performance remained unchanged. In terms of aggregation measures, all of the measures performed comparably for LDA (94-95%). PD accuracy values were also similar across the aggregation methods, but the SSD's were subject to greater variability. Bayes had the lowest SSD in both cases, with Adjusted Bayes being a close second.

Table 6.1: Disease categorization accuracy for simulated data

Method	Classifier	Average	WMLO	Adjusted Bayes	Bayes
Standard	LDA	0.95 ± 0.05	0.95 ± 0.05	0.95 ± 0.05	0.94 ± 0.05
	PD	0.96 ± 0.12	0.96 ± 0.11	0.95 ± 0.10	0.95 ± 0.10
Hierarchical	LDA	0.95 ± 0.04	0.96 ± 0.05	0.96 ± 0.06	0.97 ± 0.05
	PD	0.94 ± 0.08	0.95 ± 0.11	0.95 ± 0.12	0.95 ± 0.11

Values expressed as mean ± SSD. Categorizations are based on 10 MUPs.

Table 6.2: Confusion matrix for standard stratification method (Bayes, LDA)

	Classified As		
True Class	Normal	Myopathic	Neurogenic
Normal	99	0	1
Myopathic	3	297	0
Neurogenic	35	2	263

Based on 100 CTRL, 300 MYO and 300 NEUR muscles. Showing 41 total errors.

Examining the confusion matrix for the standard stratification method reveals the source of most of the errors made (see Table **6.2**). Clearly, many of the muscles from the neurogenic group were misclassified as being normal (false negative). In contrast, a similar type of error (false positive) was only made 3 out of 300 times for the myopathic group. Similar trends were observed in confusion matrices corresponding to all classifier-aggregation method combinations. To better understand the sources of these errors, it is useful to examine the accuracy (sensitivity) per muscle group (since each muscle group represents a specific LOI). Looking at the right-hand side of Figure 6.1 to Figure 6.4,

the per-group muscle accuracy is shown for standard disease-based stratification for the BR and AR aggregation methods for both LDA and PD. In the LDA case, the largest performance drop is seen in the 25% neurogenic group (NEUR-25), while for PD, a performance drop occurs in the normal (CTRL) group. This indicates that PD is more sensitive at detecting disease, but has poor specificity. LDA, on the other hand, has higher specificity and slightly poorer sensitivity. However, because this loss of sensitivity is localized to the NEUR-25 group, its impact is less significant when calculating SSD, thus LDA is better balanced with respect to both sensitivity and specificity.

Although total accuracy and SSD are fairly high, the notable drop in either the CTRL group or the NEUR-25 group is undesirable as it demonstrates a clear trade-off between sensitivity and specificity. The hierarchical classifier was constructed for this purpose. Table 6.3 shows the confusion matrix produced when high-resolution (LOI) stratification is used at the level of classification. Although the per-class error rate is quite high, the errors are generally confined to the same disease group. For example, although the NEUR-25 group was misclassified as being NEUR-50 or NEUR-75 29 and 5 times, respectively, it was only misclassified as normal (false negative) seven times. In total, only 11 such false negatives were made across the neurogenic LOI groups. This is a substantial improvement over the 35 false negatives seen when using standard disease-based stratification, and is summarized by the confusion matrix for the hierarchical classifier (see Table 6.4). Thus, the hierarchical classifier, constructed on this basis, was able to improve performance.

Table 6.3: Confusion matrix for high resolution (LOI) stratification (Bayes, LDA)

	Classified As						
True Class	CTRL	MYO-25	MYO-50	MYO-75	NEU-25	NEU-50	NEU-75
CTRL	88	0	0	0	10	2	0
MYO-25	0	50	36	14	0	0	0
MYO-50	0	41	30	29	0	0	0
MYO-75	0	13	15	72	0	0	0
NEU-25	7	0	0	0	59	29	5
NEU-50	3	0	0	0	24	53	20
NEU-75	1	0	0	0	10	21	68

Based on 100 muscles per category. Showing 23 total errors for disease categorization. The shaded regions represent the category breakdown within each high-level disease state.

Table 6.4: Confusion matrix for hierarchical stratification method (Bayes, LDA)

	Classified As		
True Class	Normal	Myopathic	Neurogenic
Normal	88	0	12
Myopathic	0	300	0
Neurogenic	11	0	289

Based on 100 CTRL, 300 MYO and 300 NEUR muscles. Showing 23 total errors.

In particular, total accuracy was improved by 1-4%, and SSD was decreased by 0.01-0.05 using LDA and 0.03-0.06 using PD. For LDA, Bayes aggregation had the lowest SSD at 0.01, and Average aggregation had the worst accuracy and SSD at 96% and 0.05, respectively. The opposite was observed for PD, where Average aggregation had the lowest SSD at 0.03. Although these results demonstrate improvement, their differences are subtle and do not illustrate the true benefit of the hierarchical approach. Examining per-group accuracy for the hierarchical classifier, depicted in the left-hand side of Figure 6.1 to Figure 6.4, provides insight into the level of 'balance' across individual LOI groups, as well as overall sensitivity and specificity. The LDA classifier using the BR aggregation method achieves the best balance and overall performance. For LDA, the AR aggregation methods still shows improvement over the standard classifier, however this improvement is not as substantial for the NEUR-25 group at on 0.82 compared with 0.93 for BR. For PD however, BR achieved very high sensitivity across all of the LOI groups, but specificity decreased by 0.03. The AR aggregation method achieved a better balance across groups when used with PD; however the mean accuracy actually decreased.

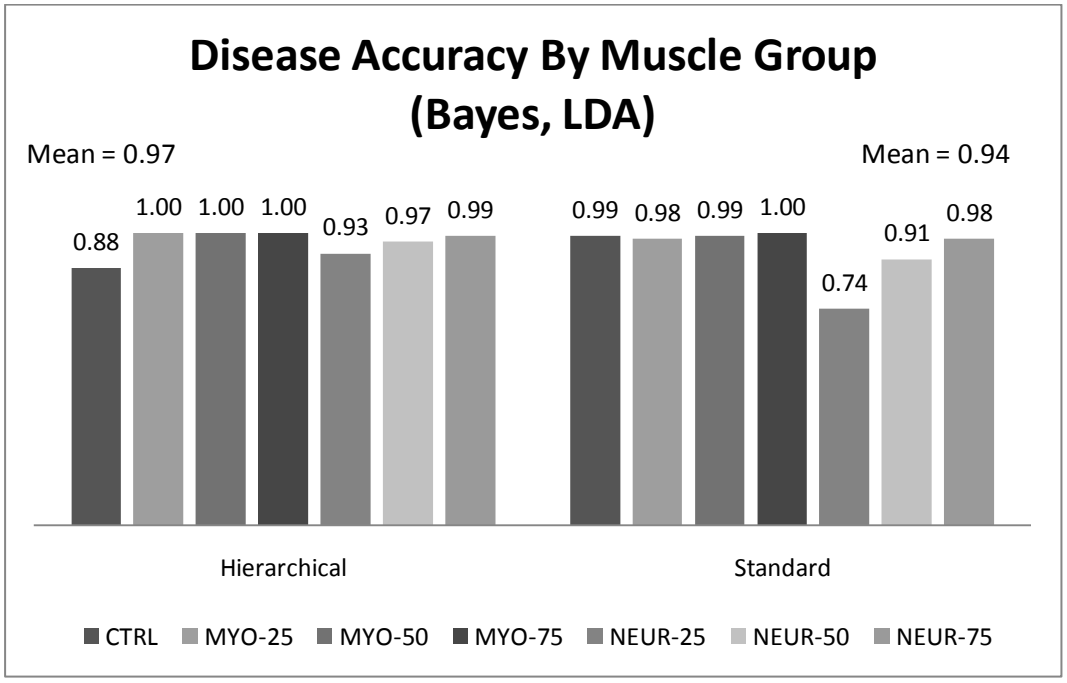

Figure 6.1: Disease categorization accuracy by muscle group (Bayes, LDA)

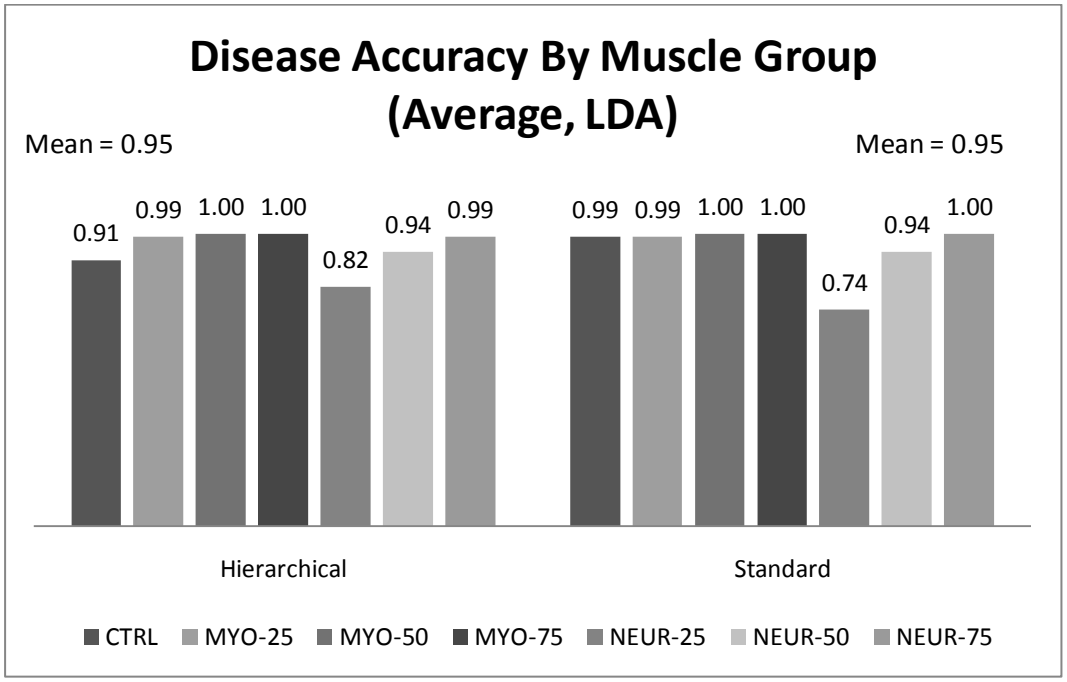

Figure 6.2: Disease categorization accuracy by muscle group (Average, LDA)

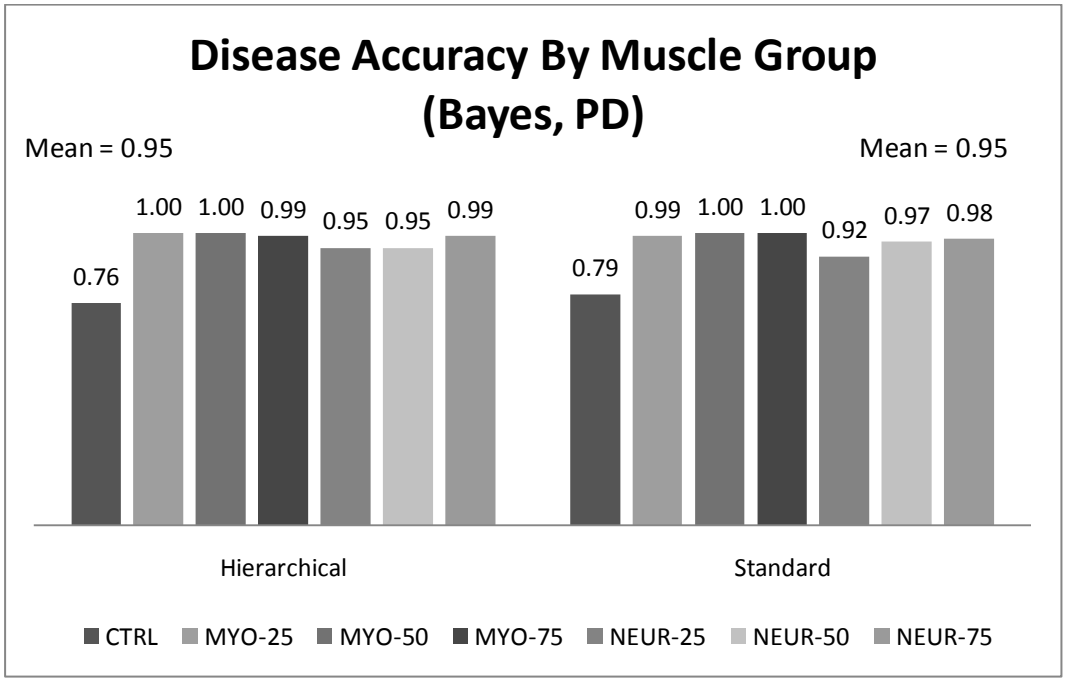

Figure 6.3: Disease categorization accuracy by muscle group (Bayes, PD)

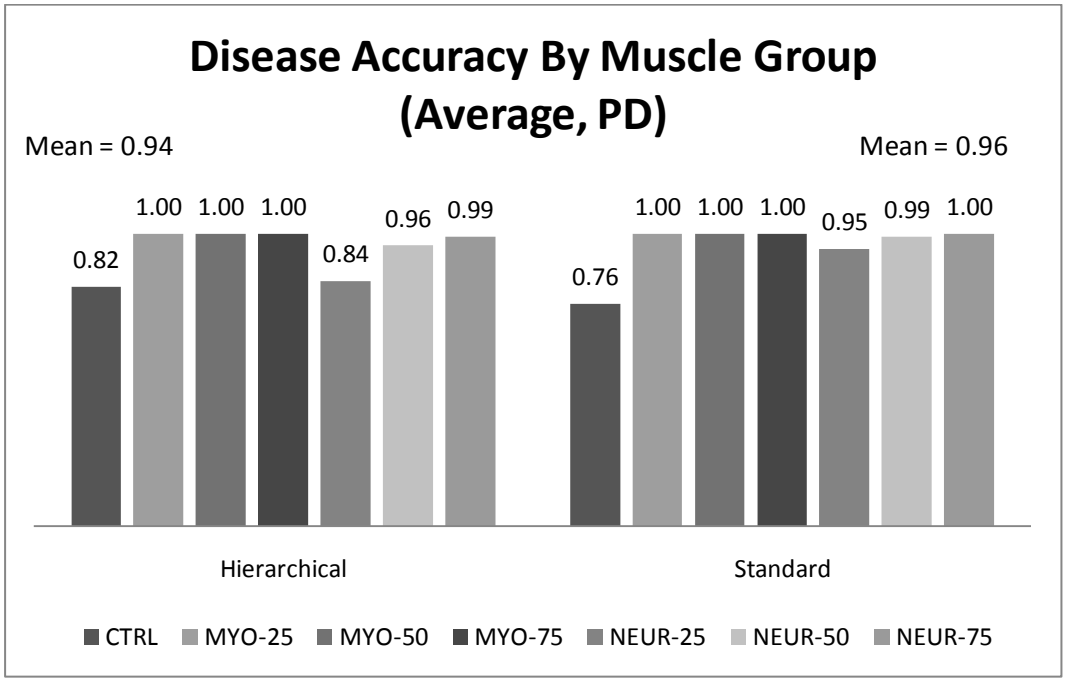

Figure 6.4: Disease categorization accuracy by muscle group (Average, PD)

6.2.2 Validation Using MDX Data

In general, total accuracy for the MDX data set was reasonable, however there was more room for improvement as compared to the simulated data set. Using disease stratification, high accuracy was achieved when muscles were comprised of either 10 or 20 MUPs. With 10 MUPs, LDA and PD were comparable with accuracies between 84-86%. LDA and PD had comparable SSD values across all aggregation measures (0.09). In fact, no change in performance was noted across the aggregation measure when PD was used, however for LDA, AR performed marginally lower than the other methods. These results are illustrated in Table **6.5**. Although not shown, an increase in performance was observed for both LDA and PD when muscles were comprised of 20 MUPs over those comprised of 10 MUPs. In all cases, increases of 3-5% were seen, and SSD values remained constant (0.08-0.09).

Table 6.5: Disease categorization accuracy for MDX Data

Method	Classifier	Average	WMLO	Adjusted Bayes	Bayes
Standard	LDA	0.85 ± 0.09	0.86 ± 0.09	0.86 ± 0.09	0.86 ± 0.09
	PD	0.84 ± 0.09	0.84 ± 0.09	0.84 ± 0.09	0.84 ± 0.09
Hierarchical	LDA	0.85 ± 0.06	0.89 ± 0.03	0.90 ± 0.01	0.91 ± 0.01
	PD	0.85 ± 0.06	0.85 ± 0.08	0.85 ± 0.09	0.85 ± 0.09

Values expressed as mean ± SSD. Categorizations are based on 10 MUPs.

Examining the confusion matrix for the standard stratification method (Table **6.6**) reveals that most of the errors made were in the disease group. While only 3 out of 100 normal muscles were misclassified as myopathic (false positive), 40 out of 200 muscles were misclassified as normal (false negative). Similar trends were observed in confusion matrices corresponding to all classifier-aggregation method combinations. To better understand the sources of these errors, it is useful to examine the accuracy (sensitivity) per muscle group (since each muscle group represents a specific LOI). Looking at the right-hand side of Figure 6.5 to Figure 6.8, the per-group muscle accuracy is shown for standard disease-based stratification for the BR and AR aggregation methods for the LDA classifier. As expected, the largest performance drop is seen in the low myopathic group (MYO-L). Accuracy for the CTRL and MYO-H groups was high because the ability to discriminate between normality and disease is improved when the disease has significantly progressed.

Table 6.6: Confusion matrix for standard stratification method - MDX data (Bayes, LDA)

	Classified As	
True Class	CTRL	MYO
CTRL	97	3
MYO	40	160

Based on 100 CTRL and 200 MYO muscles. Showing 43 errors.

Although the SSD is rather low, it is clear from these plots that disease categorization is not well balanced across LOI groups. Just as with the simulated data set, there is a clear trade-off between sensitivity and specificity.

Table **6**.**7** shows the confusion matrix produced when high-resolution (LOI) stratification is used at the level of classification. Although the per-class error rate is quite high (but not as high as it was for simulated data), the errors are generally confined to the same disease group. There was a significant drop in the number of false negative errors (17 in total) compared with the low-resolution stratification method. However, this improvement came at a cost of seven additional errors in the CTRL group (false positive). In total, 27 errors were made which is a substantial improvement over the 43 errors made when using standard disease-based stratification, as seen in the confusion matrix for the hierarchical classifier (see Table **6**.**8**). Thus, the hierarchical classifier, constructed on this basis, was able to improve performance.

In particular, total accuracy was improved by 1-5%, and SSD was decreased by 0.03-0.08 using LDA. A small increase of only 1% and a decrease in SSD of 0.01-0.03 was observed using PD. For LDA, BR aggregation had the highest accuracy (0.91) and lowest SSD at 0.01, and AR aggregation had the worst accuracy and SSD at 0.85 and 0.09, respectively. The opposite was observed for PD, where Average aggregation had the lowest SSD at 0.06, although total accuracy was unchanged across the aggregation methods. With the exception of the BR-LDA combination, these results only demonstrate a marginal improvement that does not illustrate the true benefit of the hierarchical approach. Examining per-group accuracy for the hierarchical classifier, depicted in the left-hand side of Figure 6.5 to Figure 6.8, provides insight into the level of 'balance' across individual LOI groups, as well as overall sensitivity and specificity. The LDA classifier using the BR aggregation method achieves the best balance and overall performance. For LDA, the AR aggregation method provides an

improvement, albeit not as substantial, for the MYO-L group at 0.62 compared with 0.83 for BR. For PD however, BR achieved very high sensitivity across the LOI groups, but specificity was very low (0.63). The AR aggregation method achieved a much better balance across groups, however the improvement was not as remarkable as it was for the BR-LDA combination. These trends were quite similar to those seen in the simulated data set.

Table 6.7: Confusion matrix for LOI stratification - MDX data (Bayes, LDA)

	Classified As		
True Class	CTRL	MYO-L	MYO-H
CTRL	90	10	0
MYO-L	17	77	6
MYO-H	0	8	92

Based on 100 muscles per category. Showing 27 total errors for disease categorization.

Table 6.8: Confusion matrix for hierarchical stratification method - MDX data (Bayes, LDA)

	Classified As	
True Class	CTRL	MYO
CTRL	90	10
MYO	17	183

Based on 100 CTRL and 200 MYO muscles. Showing 27 errors.

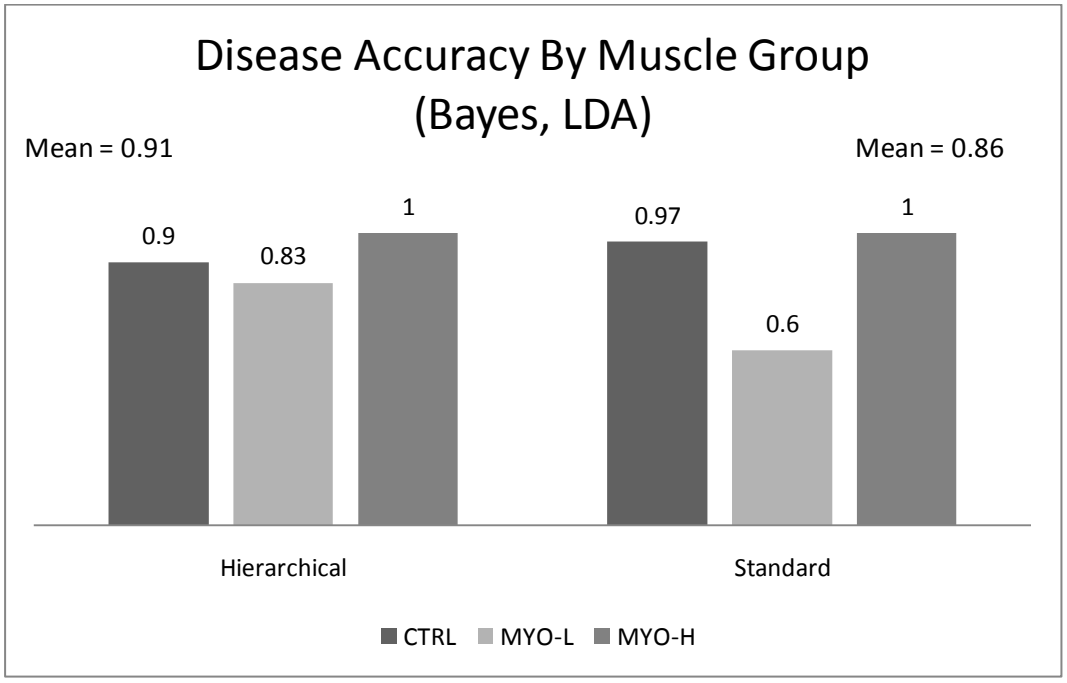

Figure 6.5: Disease categorization accuracy by muscle group - MDX Data (Bayes, LDA)

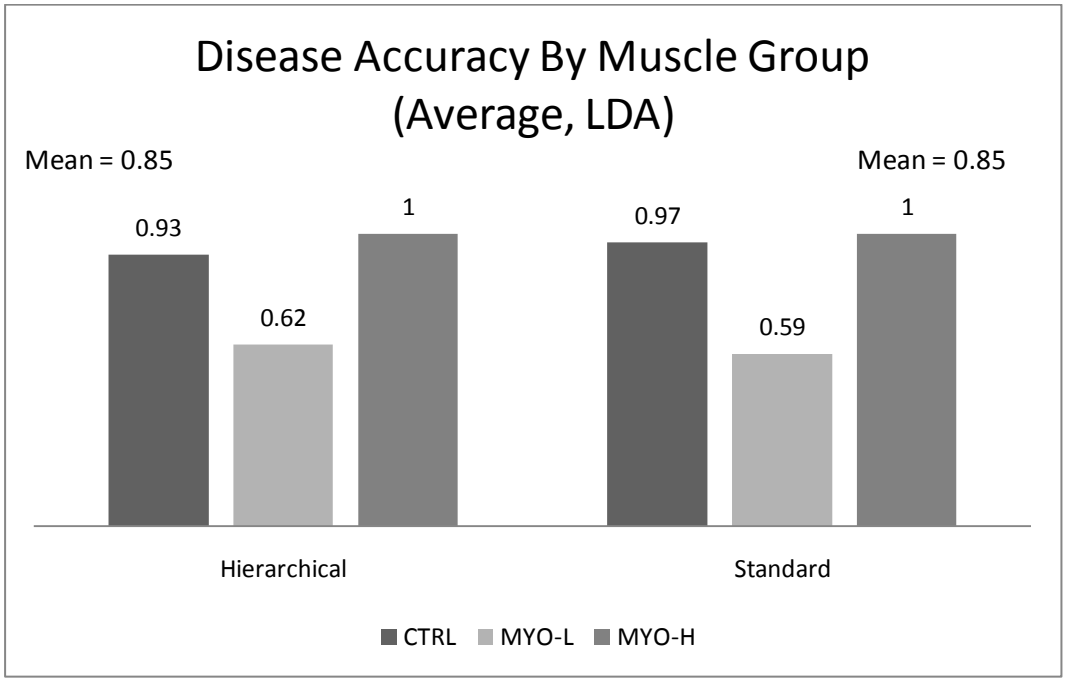

Figure 6.6: Disease categorization accuracy by muscle group - MDX Data (Average, LDA)

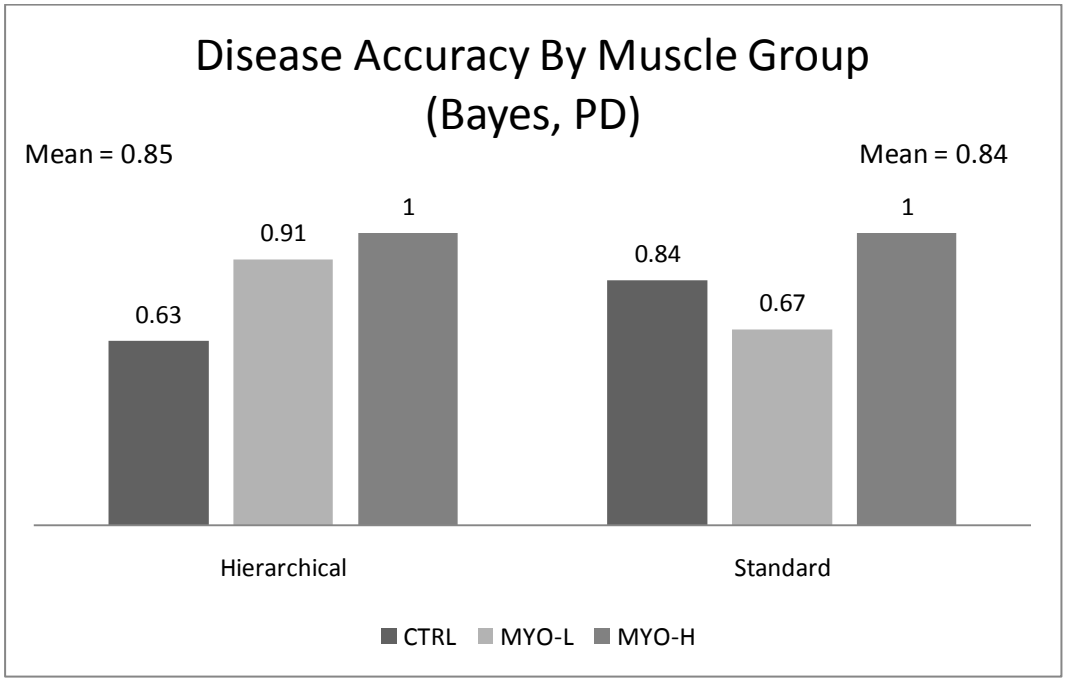

Figure 6.7: Disease categorization accuracy by muscle group - MDX Data (Bayes, PD)

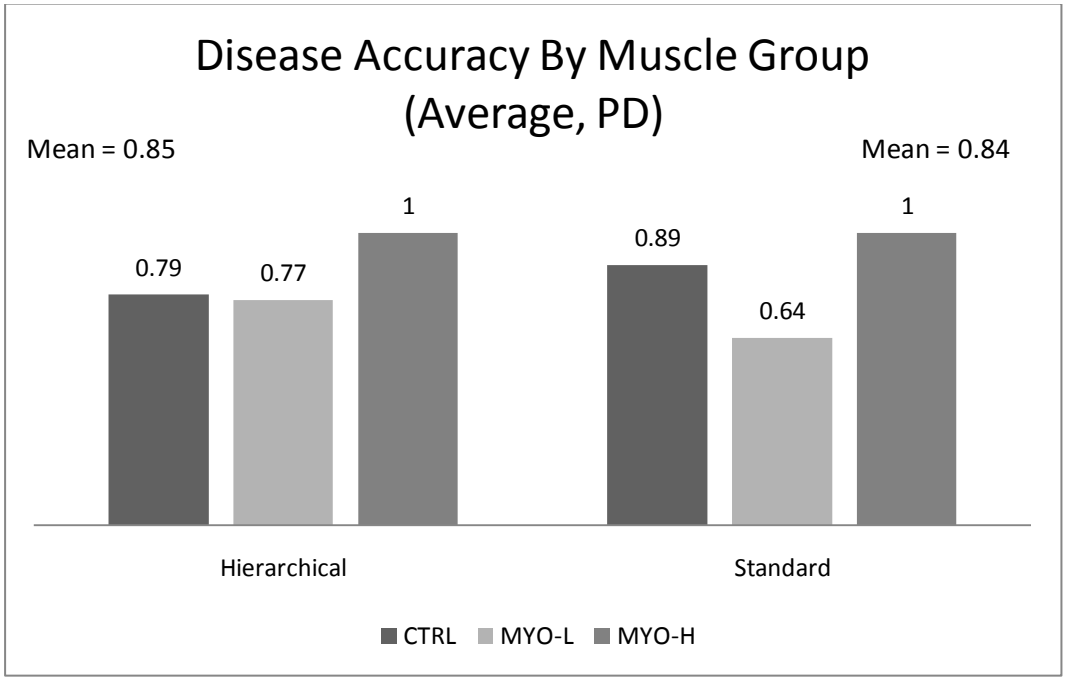

Figure 6.8: Disease categorization accuracy by muscle group - MDX Data (Average, PD)

6.3 Accuracy of LOI Categorization

The accuracy with which individual LOI categories can be detected is considered here. Various stratification methods are applied to determine if an optimal combination of class grouping and classification method yields adequate performance.

6.3.1 Evaluation Using Simulated Data

As mentioned briefly in the previous section, the categorization accuracy for LOI-group stratification was quite low. In particular, total accuracy was quite poor, with values ranging from 57-60% when using 10 MUPs and 57-64% when using 20 MUPs for LDA. Similar values were recorded for PD, ranging from 54-60%. SSD values were large in all cases, ranging from 0.12-0.26. Muscles with 10 MUPs classified by PD had the lowest SSD values (0.12-0.13) except for Average aggregation (0.17), which also had the highest SSD for all methods. When 20 MUPs were used with PD, only a slight rise in SSD (0.18) was noted, however this is of little significance in light of the poor performance overall. A general trend of a slight increase in accuracy with more evidence (more MUPs) was observed. The last row of Table 6.9 summarizes categorization accuracy for each classifier and aggregation method combination.

The two-stage classifier only had marginally better performance than the standard LOI-group stratification method. Since the error rate between myopathic and neurogenic groups was low, this was an expected finding (see Table **6**.**7**). However, it is interesting to see that SSD measures actually increased substantially from 0.12-0.20 to 0.20-0.24. The increase was greater for PD than for LDA, but was nonetheless significant for both. Despite these increases, the two-stage method was still used when applying the various LOI stratification schemes because it greatly reduces classification complexity when the number of LOI groups is high (greater than 2). A comparison of these two stratification methods is provided in the last two rows of Table **6**.**10**.

Re-examining the confusion matrix in Table **6**.**7** in terms of LOI accuracy, it is clear that applying the hierarchical method to any two adjacent LOI groups would result in better accuracy. However, in doing so, the number of clinical LOI states to report would be reduced, and the error rates among these is still substantial. Nonetheless, from a visual inspection of the errors, it is suggested that combining the 25- and 50- percent LOI groups would be optimal. In fact, the highest total accuracy was obtained when the true labels for the low and medium LOI groups were combined to form a single class. The optimal strategy was to re-assign muscles classified as medium LOI to either the low

or high LOI categories, depending on which category had the second-highest characterization score. For this strategy, the re-assignment was performed at the level of inference (and after classification). When the 25- and 50- percent LOI groups were combined prior to classification (i.e. used to re-train the classifier), a performance improvement was also observed, although to a lesser degree as compared to the re-labeling-only method.

For this optimal method, accuracy was in the range of 71-77% for LDA and 67-74% for PD. For both classifiers, SSD values were between 0.10 and 0.18 for WMLO and the Bayes-based aggregation methods, however Average aggregation had both the lowest accuracy and highest SSD (67%, 0.25). When controls were not included in classifier training, LDA accuracy fell by 2-6% for the more extreme aggregators, but increased by 3% for Average aggregation. Average aggregation SSD also dropped to 0.14, while SSD for the other methods remained the same. PD accuracy was quite poor for all methods except Average aggregation (58% vs. 64%), but even this was significantly lower that the accuracy obtained with the control group included. It should be noted however, that Average aggregation had much more consistent performance across all of the methods tested.

When the true labels of the medium and high LOI groups were combined to represent a single class, accuracy was still relatively high (71%) but SSD measures increased to 0.23. When the combined class structure was used to train the second stage classifier, accuracy was slightly poorer than re-labeling alone, with values in the range of 66-71%. SSD was also higher. The worst performance was obtained when the highest resolution of class stratification was used (i.e., individual LOI groups) regardless of whether the control group was included or not. PD consistently had slightly lower performance (lower accuracy and higher SSD) regardless of the method used.

While the exhaustive set of stratification combinations was analyzed, only the two best methods are shown in Table 6.9, along with the two-stage and standard methods for comparison.

Table 6.9: LOI categorization accuracy of optimal methods - simulated data (Bayes, LDA)

Method	Classifier	Average	WMLO	Adj. Bayes	Bayes
Re-label Only	LDA	0.71 ± 0.17	0.75 ± 0.14	0.76 ± 0.12	0.77 ± 0.12
(CTRL/LM/H)	PD	0.67 ± 0.18	0.73 ± 0.10	0.74 ± 0.10	0.73 ± 0.10
Re-Train	LDA	0.66 ± 0.20	0.68 ± 0.18	0.70 ± 0.16	0.71 ± 0.15
(CTRL/LM/H)	PD	0.59 ± 0.21	0.66 ± 0.15	0.68 ± 0.13	0.69 ± 0.13
Two-Stage	LDA	0.58 ± 0.24	0.60 ± 0.23	0.60 ± 0.22	0.61 ± 0.22
(4-class)	PD	0.55 ± 0.24	0.56 ± 0.20	0.57 ± 0.20	0.57 ± 0.20
(CTRL/L/M/H)					
Standard	LDA	0.57 ± 0.20	0.60 ± 0.17	0.60 ± 0.16	0.60 ± 0.17
(7-class)	PD	0.54 ± 0.17	0.55 ± 0.13	0.54 ± 0.12	0.54 ± 0.13
(CTRL/L/M/H)					

Values expressed as mean ± SSD. Categorizations are based on 10 MUPs.

Figure 6.9 and Figure 6.10 illustrate per-group categorization accuracy for the optimal stratification strategy compared with the two-stage LOI classifier. Notable improvements are observed, and the ability to categorize low and high LOI states is well balanced across most muscle groups, especially when using the LDA classifier.

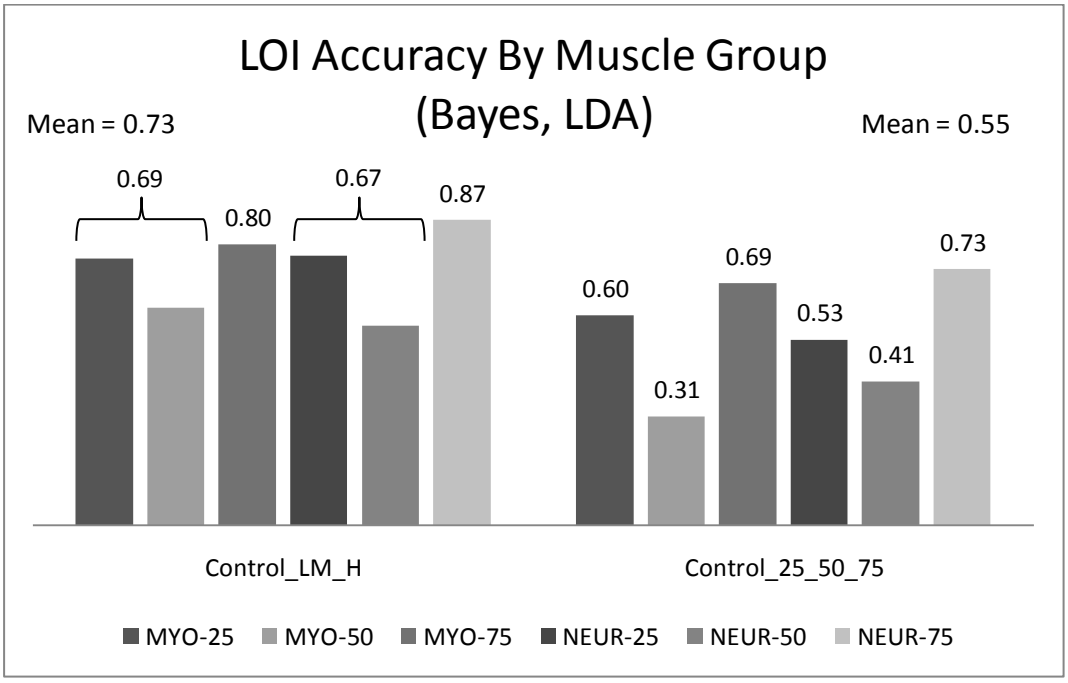

Figure 6.9: LOI categorization accuracy by muscle group - simulated data (Bayes, LDA)

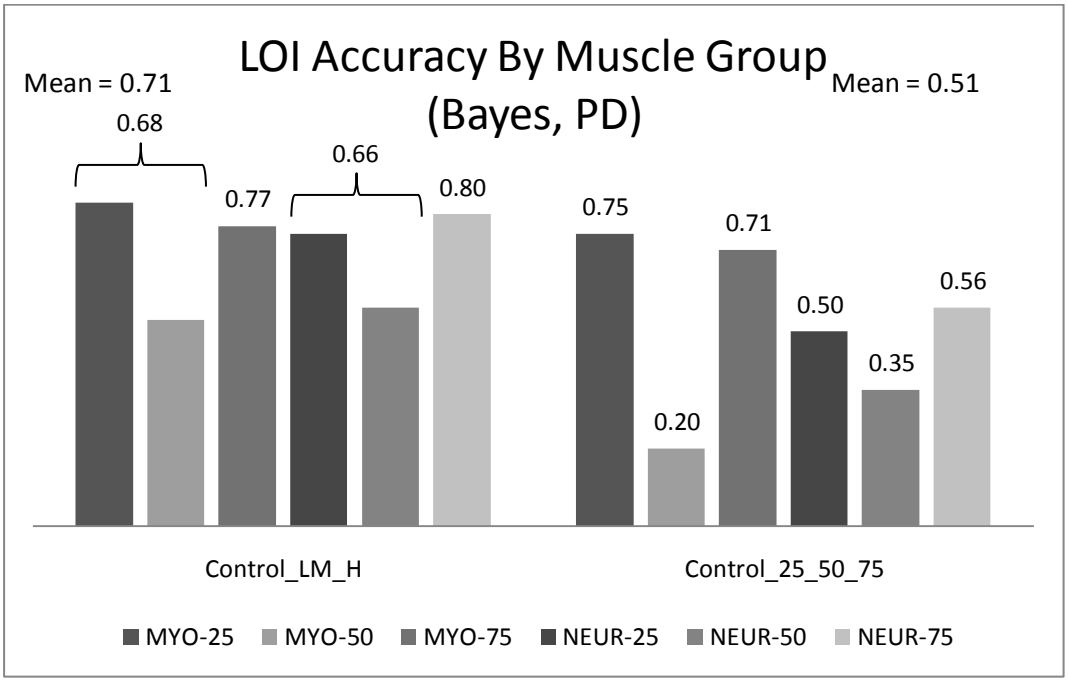

Figure 6.10: LOI categorization accuracy by muscle group - simulated data (Bayes, PD)

6.3.2 Validation Using MDX Data

The total accuracy of the high-resolution (LOI group) stratification method was much higher than that of the simulated data set. Specifically, total accuracy was better for LDA, especially with 20 MUPs, with Bayes and Adjusted Bayes aggregation reaching maximum accuracy and lowest SSD values (93%, 0.02). PD had an accuracy of 87-88% with SSD values of 0.08-0.09. For LDA, Average aggregation had the worst performance at 84% and an SSD of 0.12; however for PD there was no difference among the aggregation methods. Per-class accuracy was lowest in the low LOI group for both LDA and PD, and increased significantly (10%) with 20 MUPs. PD also had lower performance in the control group (72-86%). Of particular interest is the fact that sensitivity in detecting muscles drawn from the low LOI group improved by 12-21% for PD and 5-15% for LDA. While the improvement for PD came at a cost of specificity (down 8-21%), this was not the case with LDA.

As described in the methods section, the hierarchical classifier could not be validated on the MDX data since it was only stratified into two LOI categories. However, it was possible to evaluate the two-stage classifier, to determine the best stratification for the second stage. The optimal stratification scheme for the second stage was to use only the low and high LOI groups in a 2-class training set when the LDA classifier was used. Accuracy using this method was 88-89% and SSD values ranged from 0.01-0.02 for LDA. For PD, the optimal stratification scheme for the second stage was to use the low and high LOI groups, as well as the CTRL group, in a 3-class training set. Accuracy using this method ranged from 83-87% but SSD was much higher (0.05-0.06). For both classifiers, AR had the best performance. This performance was only marginally poorer than when disease-based stratification was used. Table **6.10** summarizes the categorization accuracy for each classifier and stratification method.

Table 6.10: LOI categorization accuracy of two-stage classifier - MDX Data (Bayes, LDA)

Method	Classifier	Average	WMLO	Adj. Bayes	Bayes
Two-Stage	LDA	0.89 ± 0.01	0.88 ± 0.02	0.88 ± 0.02	0.88 ± 0.02
(L/H)	PD	0.84 ± 0.06	0.83 ± 0.07	0.83 ± 0.07	0.83 ± 0.07
Two-Stage	LDA	0.83 ± 0.09	0.87 ± 0.05	0.87 ± 0.04	0.86 ± 0.03
(CTRL/L/H)	PD	0.86 ± 0.06	0.87 ± 0.05	0.85 ± 0.06	0.83 ± 0.06
Two-Stage	LDA	0.88 ± 0.06	0.89 ± 0.05	0.89 ± 0.05	0.89 ± 0.05
(CTRL/MYO)	PD	0.85 ± 0.08	0.87 ± 0.06	0.87 ± 0.06	0.87 ± 0.06

Values expressed as mean ± SSD. Categorizations are based on 10 MUPs.

Figure 6.11 and Figure 6.12 show a comparison of per-group categorization accuracy for the two optimal stratification strategies used in the two-stage LOI classifier. In Figure 6.11, notable improvements in balance across muscle groups are observed for LDA using only the L and H categories, while in Figure 6.12, improvements in balance across muscle groups are seen using the CTRL, L and H categories.

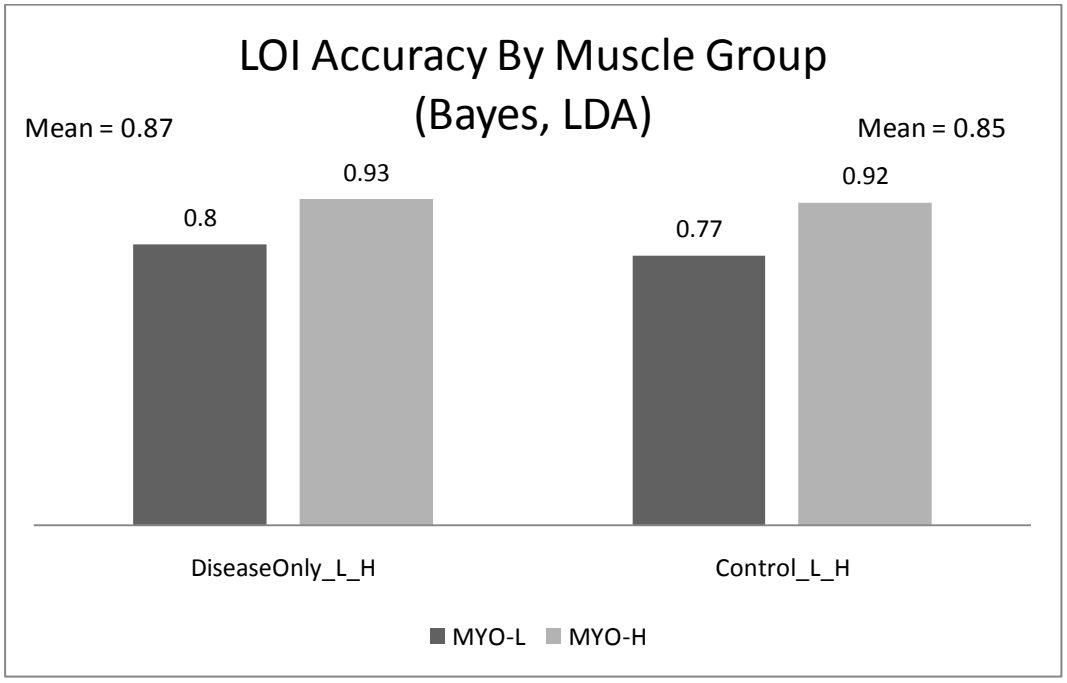

Figure 6.11: LOI categorization accuracy by muscle group - MDX data (Bayes, LDA)

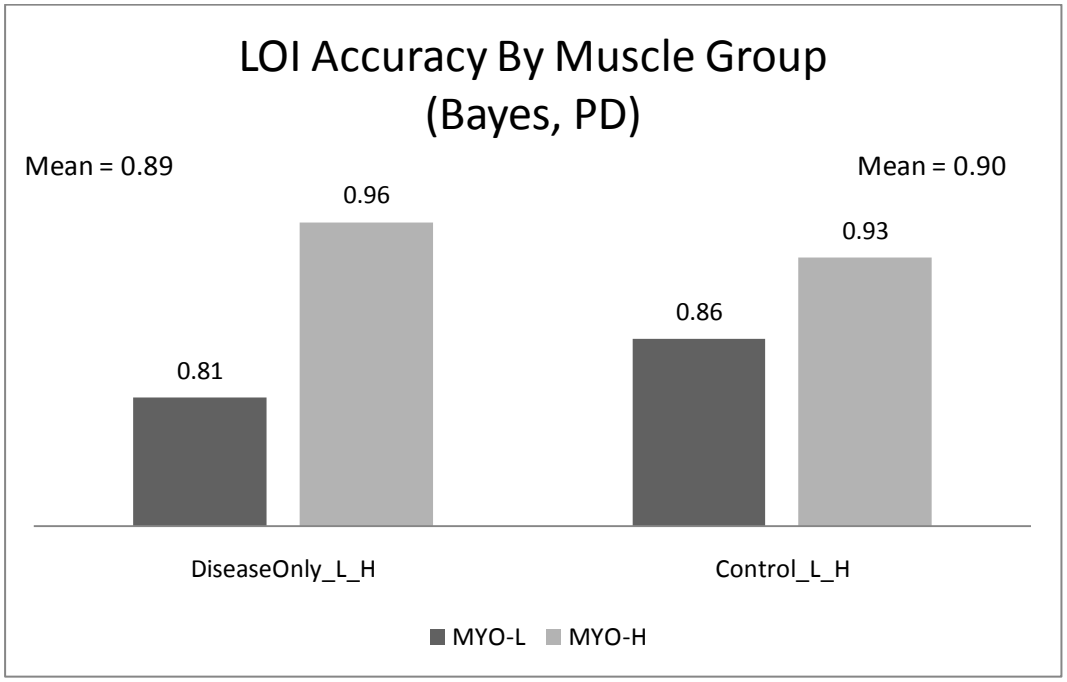

Figure 6.12: LOI categorization accuracy by muscle group - MDX data (Bayes, PD)

6.4 Continuous Measures of LOI

Previous works by Pino showed that measures of confidence correlate well with LOI when using simulated data (Pino et al., 2008). As such, similar correlation studies were performed here using Spearman's Ranking Correlation Coefficient.

6.4.1 Evaluation Using Simulated Data

The muscle characterization score relating to confidence was first assessed. When 10 MUPs were used, correlation values were quite poor in general; however the neurogenic category using the LDA classifier was significantly better than the other methods with a value of 0.62. This value was obtained when the data was stratified by disease category. Using the other stratification methods resulted in similar values. Myopathic correlation values were extremely low, and often negative. No stratification method was significantly better than any of the others. PD correlation scores were consistently poorer than LDA correlation. As the number of MUPs was increased, correlation improved for the neurogenic category, but remained the same for the myopathic category. Figure 6.13 illustrated this increasing trend for the neurogenic case.

When discrete LOI was considered, an interpolation scheme was used to compute a scalar measure that was believed to relate to LOI. This measure resulted in a significant increase in correlation, with a maximum value of 0.59 for the myopathic category, and 0.74 for the neurogenic category. For both categories, this optimal correlation was produced using the Average aggregation method. This high correlation was realized using training data that was stratified by LOI group, without including the control group. When the low and medium classes were combined prior to training, correlation was only slightly reduced, giving the second highest performance. These results are summarized in Table 6.11.

Figure 6.14 to Figure 6.17 illustrate correlation by plotting muscle scores against of the true LOI category. For the neurogenic category, and using the LDA classifier with AB or AR aggregation methods (Figure 6.14 and Figure 6.15), the tendency is for muscle scores to increase with LOI. A key difference between these methods, however, is that the AB scores are much closer to 1 in all cases while the AR scores occupy a much lower range. While many of the AB scores saturate, the saturation effect is not as bad as with BR (see Figure 6.16). WMLO produced a fairly diverse spread of values, but was not correlated well. Figure 6.17 illustrates muscle scores produced by the interpolation scheme mentioned previously. These scores are not true conditional probabilities, and

only represent a deterministic mapping. However, they score reasonably well in terms of correlation as indicated in the plot (see figure).

In general, the correlation plots reveal that there is considerable variation in muscle score, resulting in significant overlap across LOI groups.

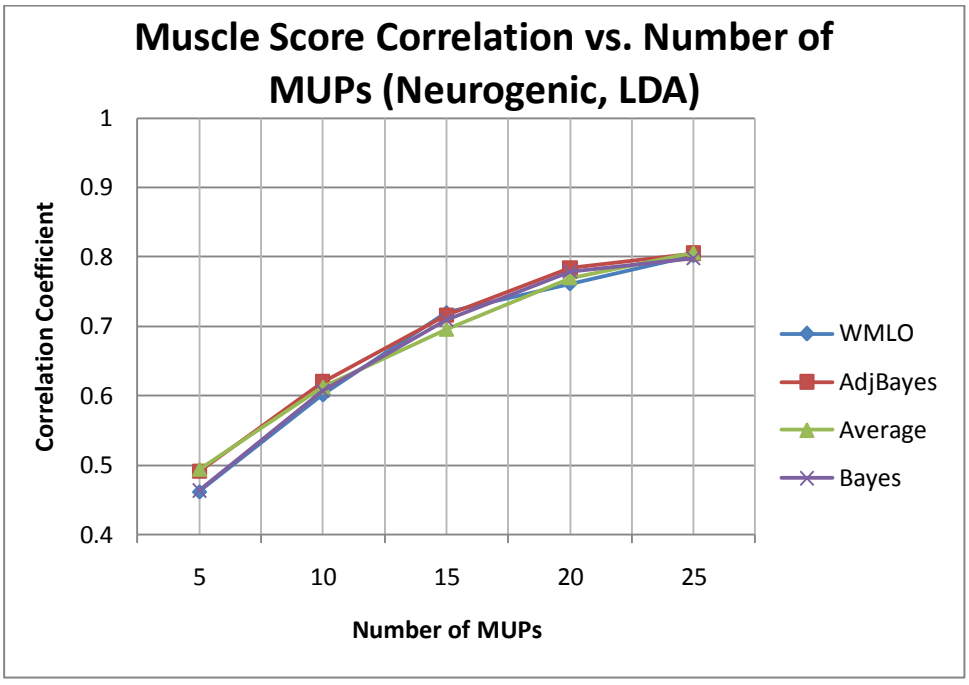

Figure 6.13: Correlation of muscle confidence scores with actual LOI – simulated data (Neurogenic, LDA)

Table 6.11: LOI correlation for best stratification methods - simulated data (Bayes, LDA)

Method	Classifier	Category	Average	WMLO	Adj. Bayes	Bayes
Re-label Only	LDA	Myopathic	0.59	0.57	0.58	0.58
(LM/H)		Neurogenic	0.74	0.71	0.72	0.71
	PD	Myopathic	0.65	0.58	0.60	0.58
		Neurogenic	0.62	0.49	0.53	0.48
Re-Train	LDA	Myopathic	0.54	0.55	0.57	0.56
(CTRL/L/M/H)		Neurogenic	0.74	0.72	0.73	0.72
	PD	Myopathic	0.48	0.53	0.54	0.55
		Neurogenic	0.64	0.57	0.60	0.57
Standard (3-class)	LDA	Myopathic	-0.13	-0.35	-0.24	-0.26
(CTRL/MYO/NEUR)		Neurogenic	0.61	0.60	0.62	0.61
	PD	Myopathic	0.00	-0.10	-0.02	-0.01
		Neurogenic	0.48	0.44	0.46	0.44

Values expressed as mean ± SSD. Categorizations are based on 10 MUPs.

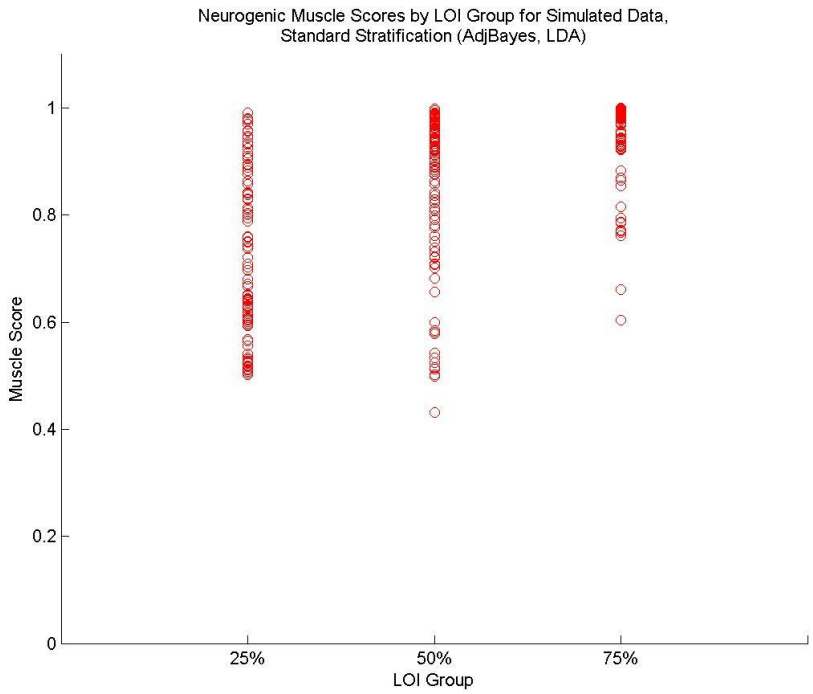

Figure 6.14: Neurogenic muscle scores by LOI group - simulated data (AB, LDA)

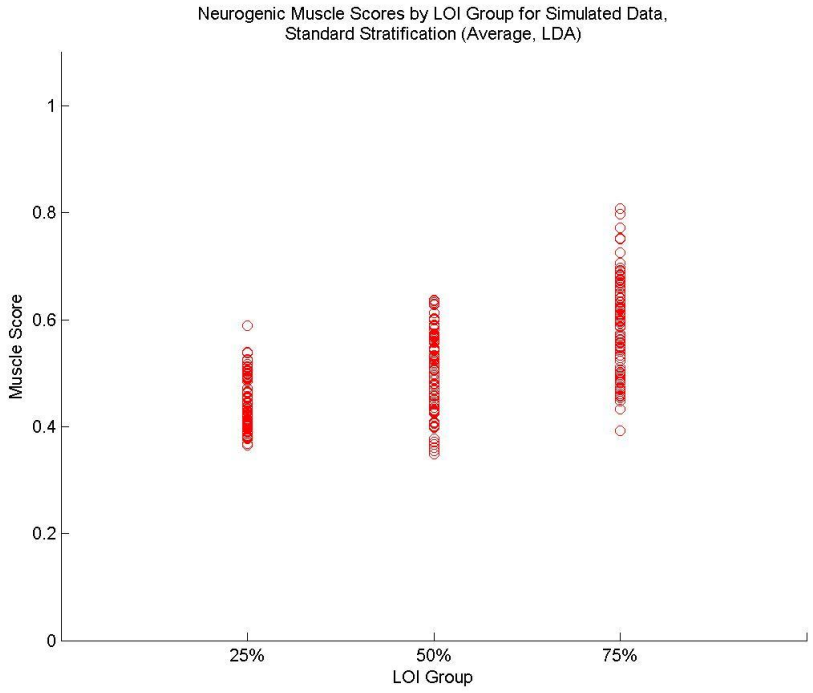

Figure 6.15: Neurogenic muscle scores by LOI group - simulated data (AR, LDA)

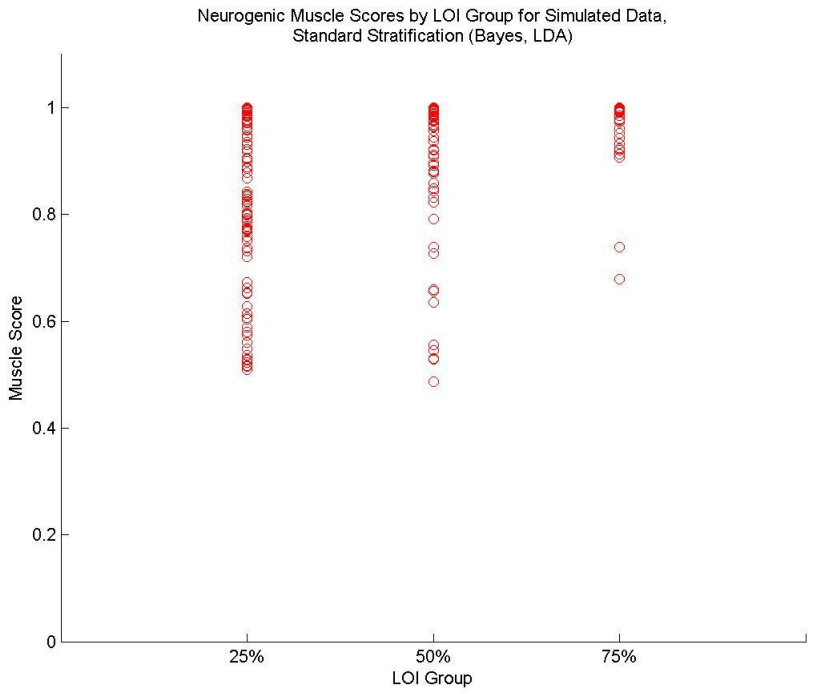

Figure 6.16: Neurogenic muscle scores by LOI group - simulated data (BR, LDA)

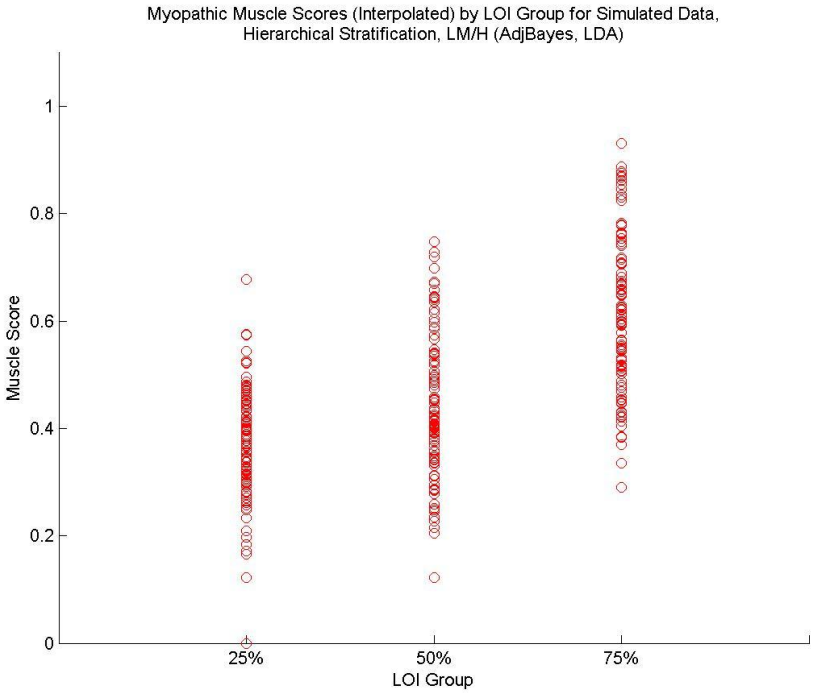

Figure 6.17: Interpolated myopathic muscle scores by LOI group - simulated data (AB, LDA)

6.4.2 Validation Using MDX Data

Correlation values produced by the MDX data set were surprisingly well correlated with LOI, unlike for the simulated data case. When 10 MUPs were used, the highest correlation was achieved by the LDA classifier and AR aggregation method combination. However, performance was quite close using the other aggregation methods. LDA again had higher correlation values than PD. As the number of MUPs was increased, correlation improved very slightly, remaining fairly constant for MUP numbers greater than 10. Figure 6.18 illustrates this trend. The best correlation was obtained when only two classes (L and H) were used to train the classifier. These results are summarized in Table 6.12.

Figure 6.19 and Figure 6.20 illustrate correlation by plotting muscle scores against of the true LOI category. When the AB method was used, muscle scores for the L group occupied the entire range of possible values, while scores for the H group were quite close to one. Using the AR aggregation method, greater separation was observed between the groups (see Figure 6.19). In fact, only a few assigned scores seem to overlap, and it seems that scores assigned to the two LOI categories are quire separable.

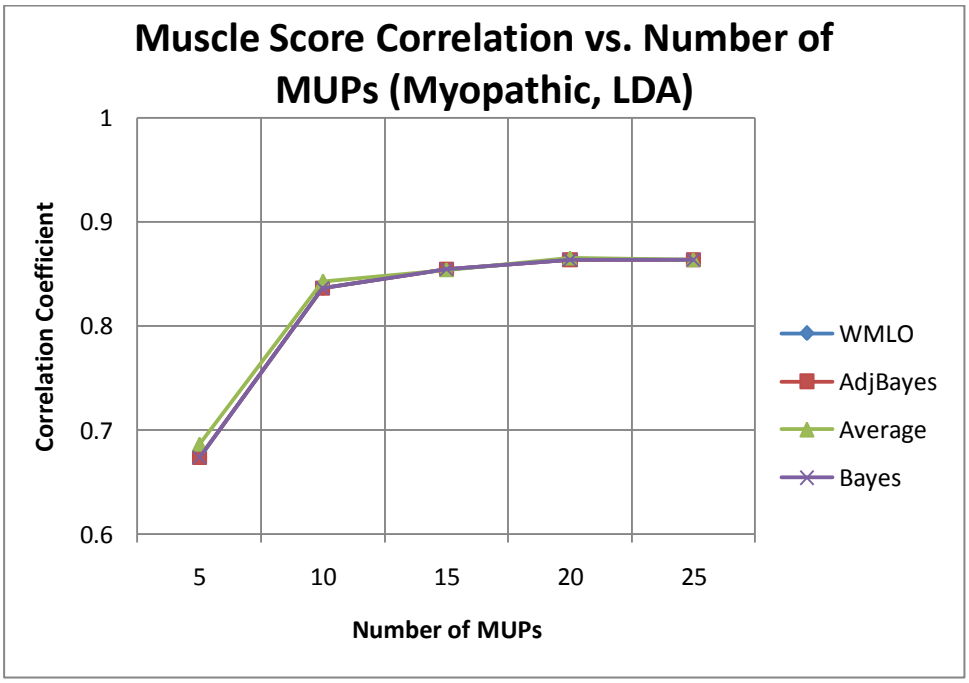

Figure 6.18: Muscle score correlation with actual LOI - MDX Data (Myopathic, LDA)

Table 6.12: LOI correlation for optimal stratification methods - MDX data (Bayes, LDA)

Method	Classifier	Average	WMLO	Adj. Bayes	Bayes
Two-Stage	LDA	0.86	0.85	0.85	0.85
(L/H)	PD	0.83	0.83	0.83	0.83
Two-Stage	LDA	0.85	0.85	0.85	0.85
(CTRL/L/H)	PD	0.83	0.83	0.83	0.82
Standard	LDA	0.82	0.82	0.82	0.82
(CTRL/MYO)	PD	0.78	0.77	0.77	0.77

Values expressed as mean ± SSD. Categorizations are based on 10 MUPs.

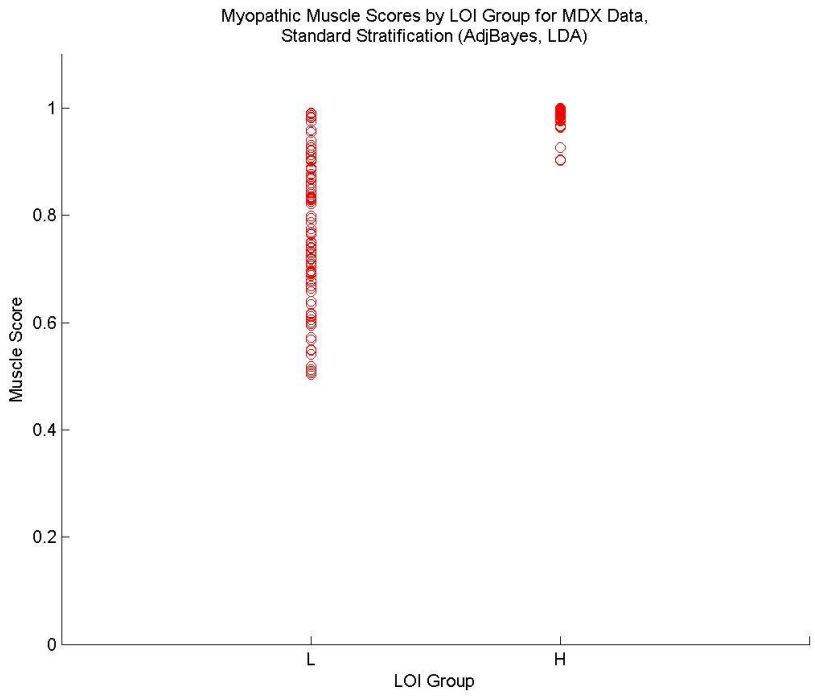

Figure 6.19: Myopathic muscle scores by LOI group - MDX data (AB, LDA)

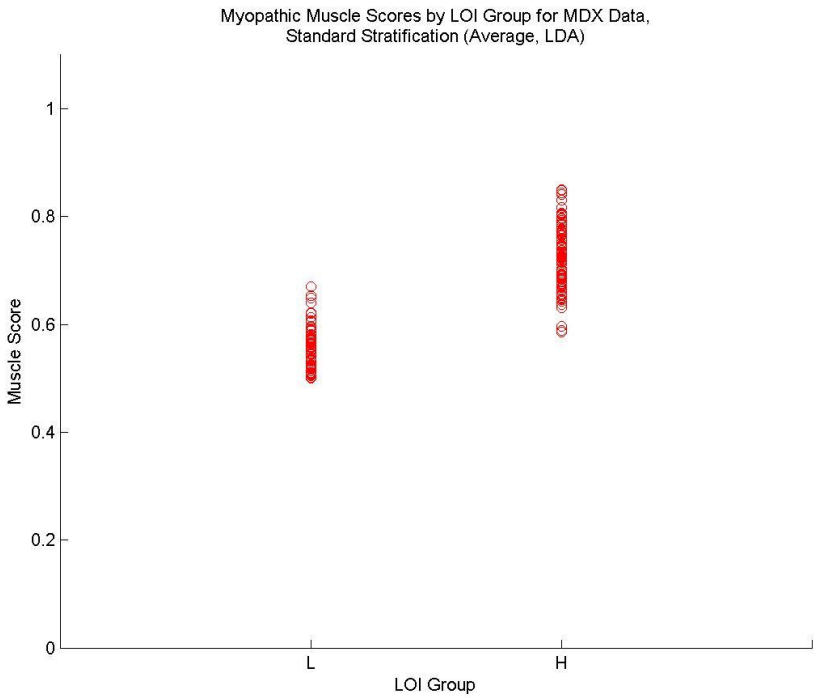

Figure 6.20: Myopathic muscle scores by LOI group - MDX data (AR, LDA)

6.5 Aggregation Measures

The accuracy of disease categorization as a function of the amount of evidence is presented next.

6.5.1 Evaluation Using Simulated Data

The results corresponding to the data used for evaluation are presented. Figure 6.21 to Figure 6.24 show categorization as a function of the number of MUPs for the LDA and PD classifiers using the standard or hierarchical stratification methods. Accuracy for each of the aggregation measures is plotted on the same graph, and represents the average accuracy of the CTRL and NEUR-25 muscle groups. Only these groups were chosen because they accentuate the differences as the other groups had accuracy values close to 1 and did not vary across methods.

When standard disease stratification was used, the aggregation measures are quite close for LDA. For PD however, BR, AB, and to a lesser extent WMLO, tend to perform better than AR. WMLO seems to reflect the average trend, lying roughly in the middle of the other methods. The accuracy of

PD as compared to LDA is slightly higher for all numbers of MUPs. However, it should be emphasized that these results are based solely on the CTRL and NEUR-25 groups, and therefore are not contradictory to the results presented in the previous section. Another notable finding is that for LDA, when 25 MUPs are used, accuracy actually drops slightly compared to when 20 MUPs are used. This is contrary to the expectation that accuracy should improve with more evidence. Also, for PD, the aggregation measures are ranked as expected at lower MUP numbers, however the trend fails at 25 MUPs. For LDA, the measures are not ranked as expected until there are at least 25 MUPs, however these differences are less significant since performance is quite comparable across each method.

When the hierarchical method is used, very different behaviors are observed across the classifiers and aggregation methods. For the LDA classifier, performance is drastically improved when the number of MUPs is greater than 10 for WMLO, AB, and BR. The performance improvement tapers off as the number of MUPs exceeds 15. Note how these measures practically overlap each other. The AR method, on the other hand, lags considerably in terms of accuracy. In fact, for higher MUP numbers, it seems to follow a similar trajectory as in the standard stratification method, including the notable dip at 25 MUPs. Despite significant overlap, it appears that the rankings of each method are as expected. For the PD classifier, the opposite effect is observed. Mainly, the AR method excels, particularly for MUP numbers greater than 15, while the other methods perform quite closely to each other. At 20 MUPs, the methods are all equivalent, however AR finishes on top at 25 MUPs. Again, the aggregation measures are ranked as expected, except at very low MUP numbers. The average performance of LDA is higher than that of PD, except when AR is used.

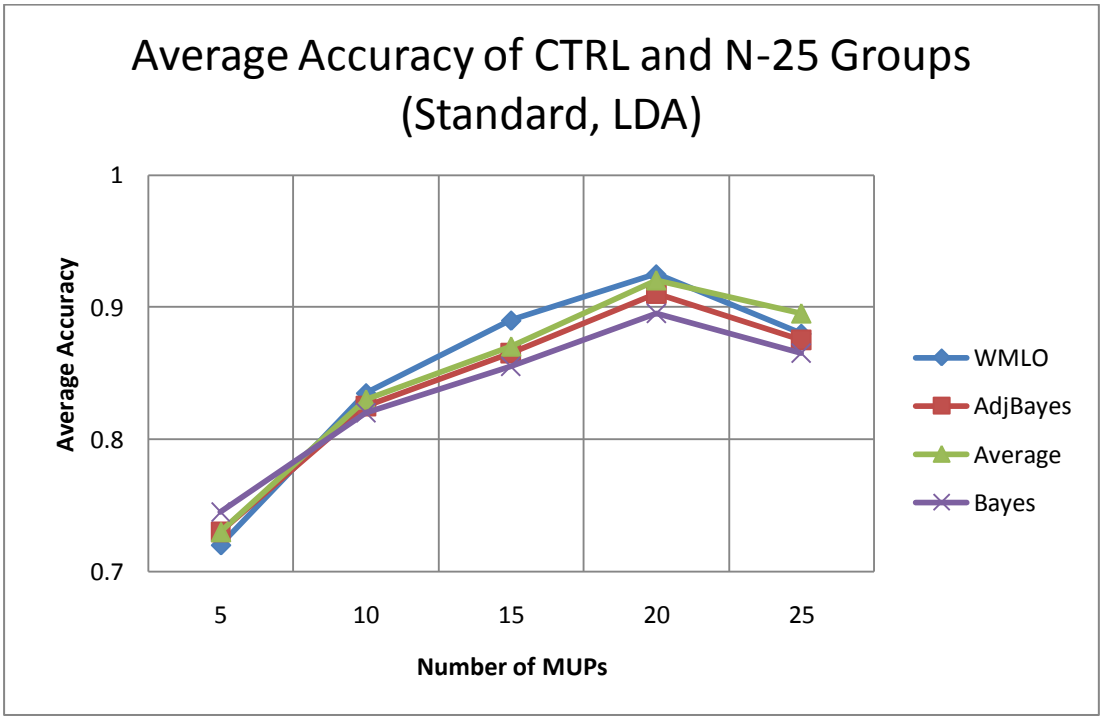

Figure 6.21: Average accuracy vs. number of MUPs – simulated data (Standard, LDA)

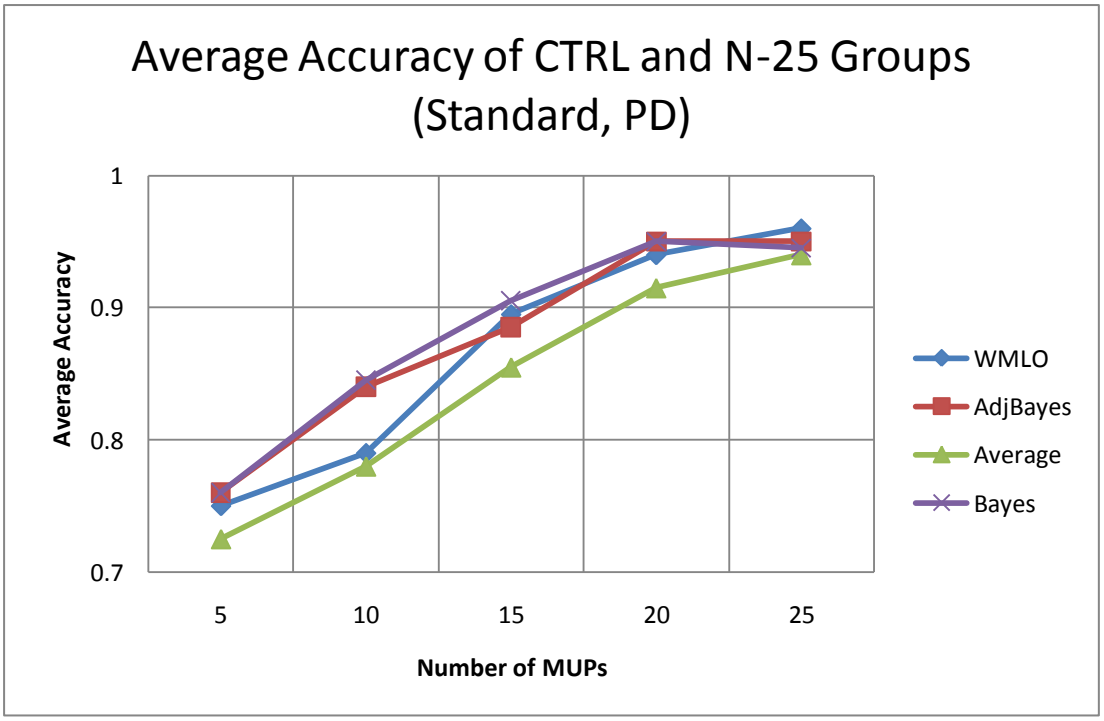

Figure 6.22: Average accuracy vs. number of MUPs – simulated data (Standard, PD)

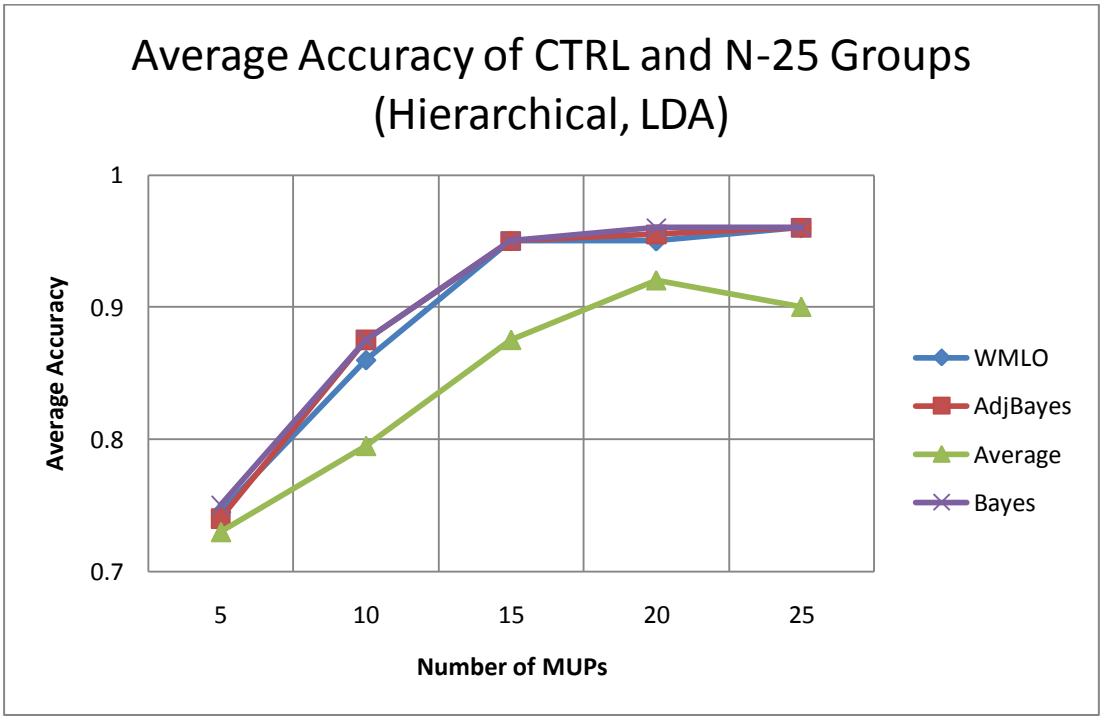

Figure 6.23: Average accuracy vs. number of MUPs – simulated data (Hierarchical, LDA)

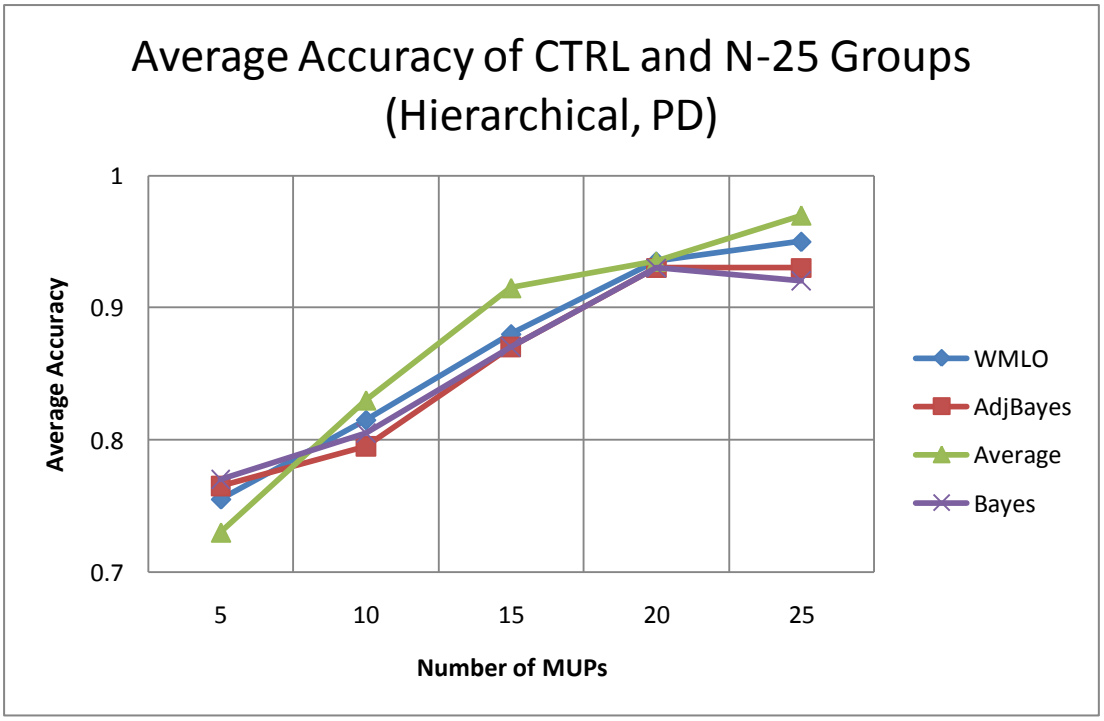

Figure 6.24: Average accuracy vs. number of MUPs – simulated data (Hierarchical, PD)

6.5.2 Validation Using MDX Data

The accuracy of disease categorization as a function of the amount of evidence is presented next. Figure 6.25 to Figure 6.28 show categorization as a function of the number of MUPs for the LDA and PD classifiers using the standard or hierarchical stratification methods. Accuracy for each of the aggregation measures is plotted on the same graph, and represents the average accuracy of the CTRL and MYO-L muscle groups. Only these groups were chosen because they accentuate the differences as the MYO-L had accuracy values close to 1 and did not vary across methods.

When standard disease stratification was used, the aggregation measures were quite close and overlapped for LDA. For PD however, BR, AB, and WMLO are completely overlapped and tend to perform better than AR. The accuracy of PD as compared to LDA is slightly lower for all numbers of MUPs except for 25 MUPs, where there is an increase. Another notable finding is that for LDA, when 25 MUPs are used, accuracy actually drops slightly compared to when 20 MUPs are used. This characteristic was also observed in the simulated data set and is contrary to the expectation that accuracy should improve with more evidence. No comments can be made about the ranking of aggregation methods because of the significant overlap.

When the hierarchical method is used, very different behaviors are again observed across the classifiers and aggregation methods. For the LDA classifier, performance is drastically improved when the number of MUPs is greater than 15 for WMLO, AB, and BR. The AR method, on the other hand, lags considerably and seems to follow a similar trajectory as with the standard stratification method, including the notable dip at 25 MUPs. The rankings of the aggregation methods are as expected, except for the fact that AB and BR tend to switch places for MUP values of 15 or more. However, given the closeness of these two methods, this change of position is not significant.

For the PD classifier, performance improves marginally as compared to the standard stratification method. However, there is an unexpected drop in performance at 25 MUPs for all of the aggregation measures. Interestingly, the rankings of each method are in the proper order, with AR showing the highest accuracy, despite the fact that this accuracy is lower than that of the standard classifier. The performance of the LDA classifier is significantly better than PD for all aggregation measures except AR.

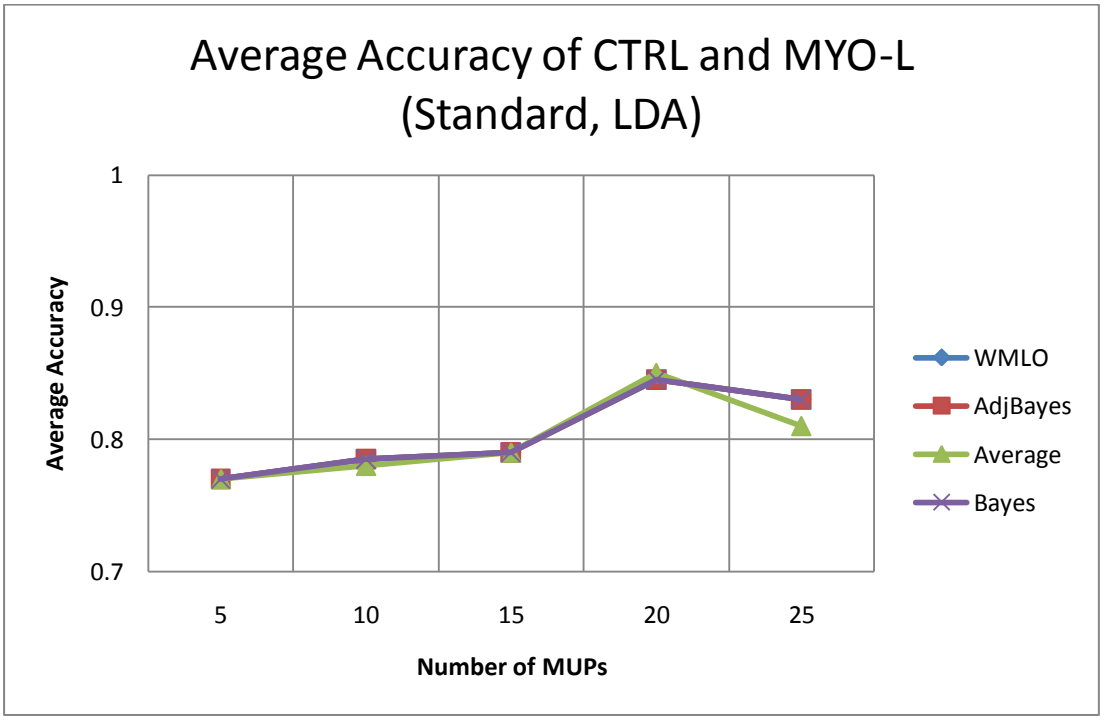

Figure 6.25: Average accuracy vs. number of MUPs – MDX data (Standard, LDA)

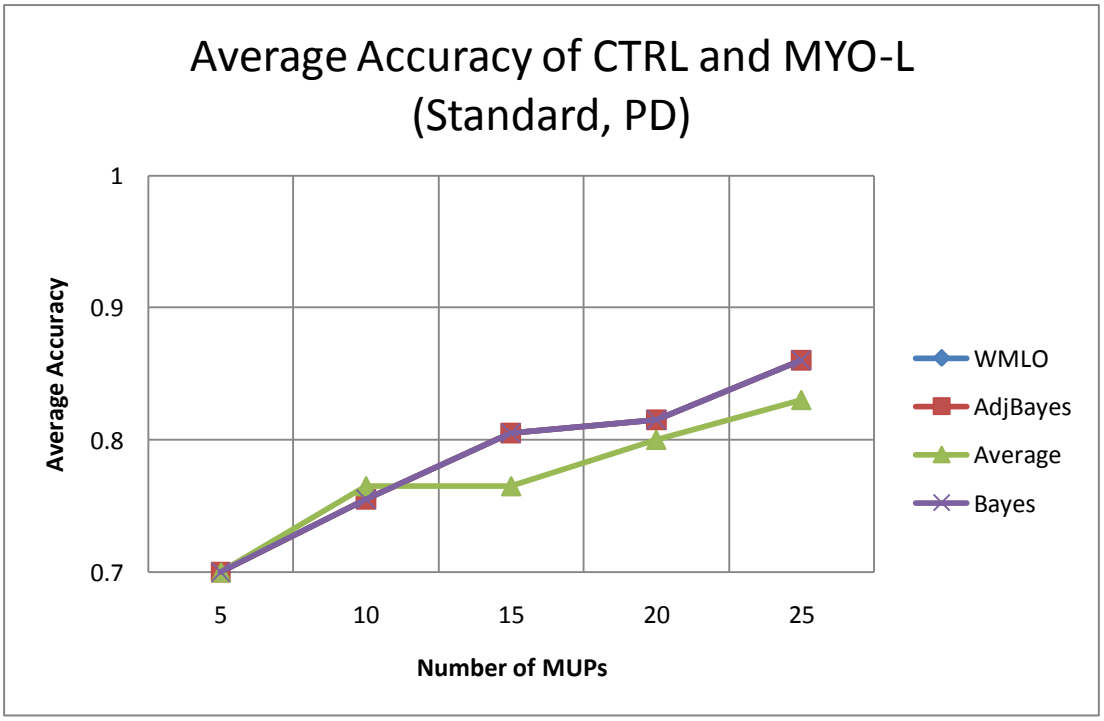

Figure 6.26: Average accuracy vs. number of MUPs – MDX data (Standard, PD)

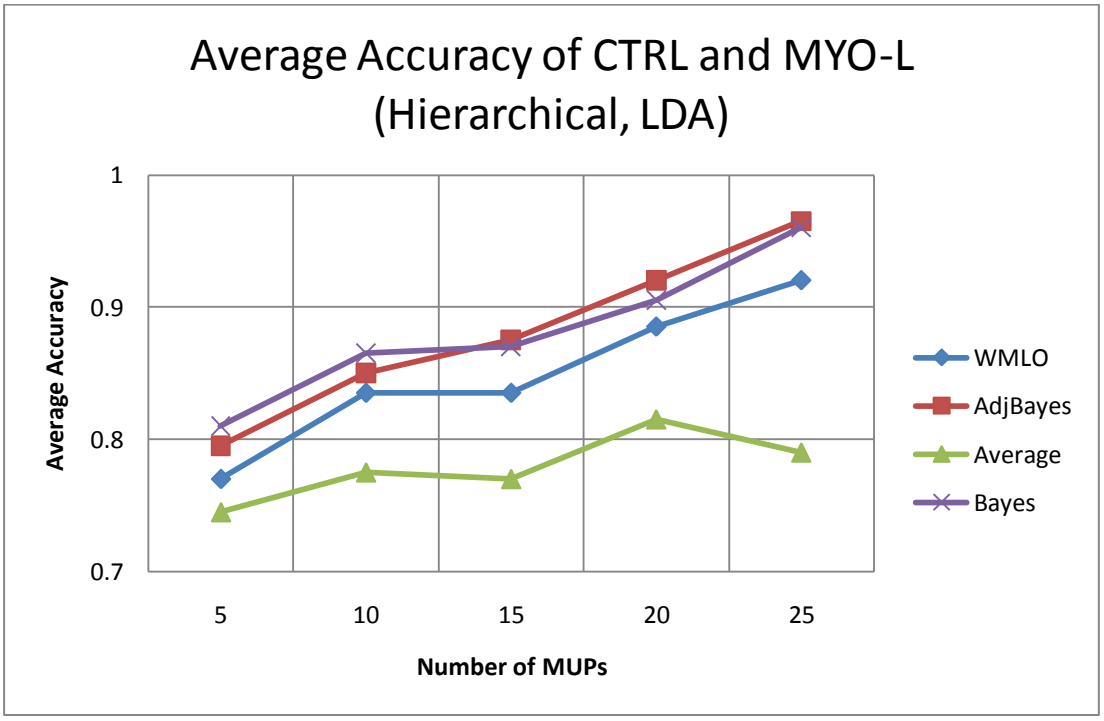

Figure 6.27: Average accuracy vs. number of MUPs – MDX data (Hierarchical, LDA)

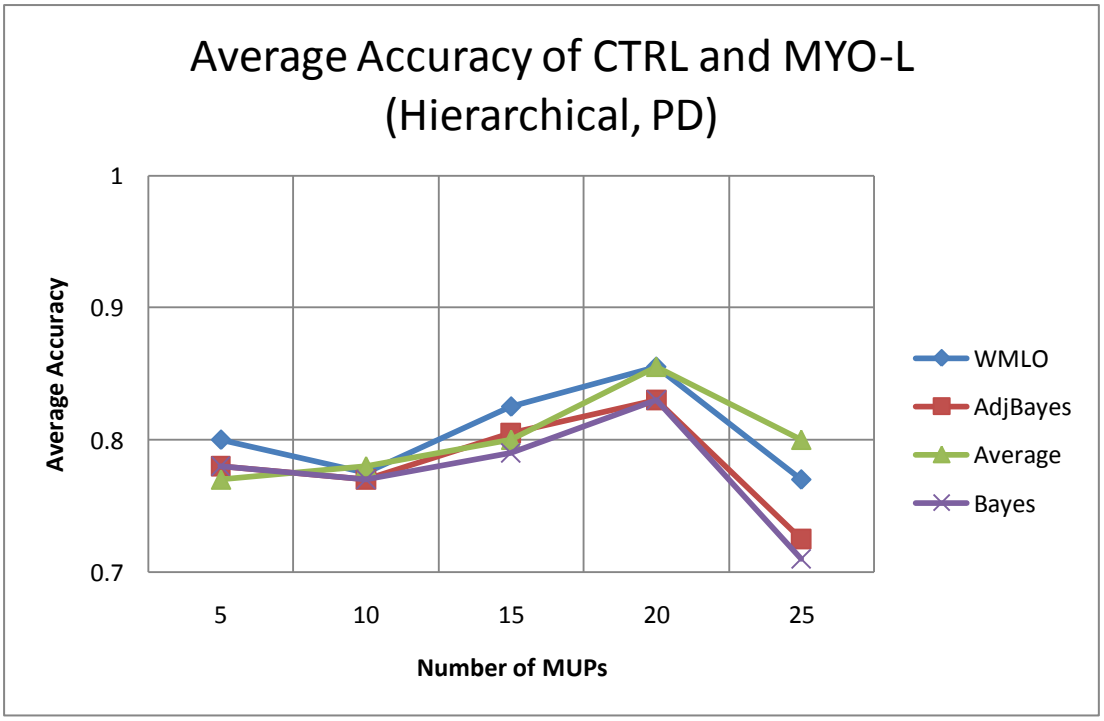

Figure 6.28: Average accuracy vs. number of MUPs – MDX data (Hierarchical, PD)

Chapter 7
Discussion

In order to improve the accuracy with which disease categories could be detected, a hierarchical stratification strategy was evaluated. The findings related to this strategy are evaluated to determine if a significant improvement was observed. Next, the various LOI strategies are discussed to identify the optimal combination that is able to provide clinicians with useful information relating to LOI, and with reasonable accuracy. Continuous measures of LOI are shown to be quite poor in general, and their efficacy in provided clinically relevant information is considered. Finally, an analysis of the behaviors of the various aggregators is able to provide valuable insight in selecting an appropriate method for use in a CDSS.

7.1 Accuracy of Disease Categorization

The accuracy of disease categorization is discussed in terms of total accuracy, as well as the accuracy across individual LOI groups. A classifier that performs reasonably well across all categories is desired.

7.1.1 Evaluation Using Simulated Data

In general, the accuracy of disease categorization was considerably high, with total accuracy above 95% in most cases. Although some changes were observed, there is very little difference in performance whether 10 or 20 MUPs are considered. Global performance (that is, performance across all muscle groups) was high for two reasons: first, because the underlying classification problem was not inherently difficult (simulated data was defined to have large increments with respect to LOI), and secondly, because poorer performance in any one of the muscle groups is compensated for by high accuracy in the other groups. Thus global performance alone is insufficient for describing the clinical performance of any particular method. An analysis of the per-class accuracies revealed that there was a particular group, specifically the NEUR-25 group, which had significantly poorer accuracy than the other groups. This is typical in the case of neurogenic disease because in the earlier stages of progression there is a lower chance of MU restructuring and some MU's may even appear myopathic with respect to certain features. The latter was not the case because there were very few errors made between disease categories. Re-examining the feature plots for the various LOI groups of the

neurogenic category reveals that there is in fact considerable overlap between the CTRL and NEUR-25 group. However, even with this overlap, fairly high accuracy was obtained.

While the performance is not inherently bad, and may be acceptable from a pattern recognition perspective, it is not clinically acceptable to have a significant drop in performance in one specific muscle group. In the area of disease detection, it is often the case that early detection leads to more effective intervention and treatment and thus a higher cost might be associated with miss-detection of a lower level of disease. However, the contrary may also be true: higher specificity may be desired over higher accuracy of lower levels of disease because a false positive might cause undue stress and mental anguish for the patient. Despite the costs, however, it is clear that a better balance between sensitivity and specificity should be sought.

The hierarchical scheme was able to provide such a balance, showing a significant increase in detection accuracy of the N-25 group, without a significant loss of accuracy in the other groups. The improvement was more significant for LDA than for PD as PD already had good performance using disease stratification. A likely explanation for this is the fact that PD can detect complex patterns in the data and thus form non-linear decision boundaries, giving it an advantage over the linear boundaries formed by LDA. By increasing the stratification resolution, the number of training samples per class is decreased. Since PD is more dependent on training set size in order to detect significant patterns, the ability for improvement when using the hierarchical classifier is subject to the availability of a sufficiently large set of training data.

Nonetheless, in the case of LDA, and to a smaller extent for PD, the hierarchical method offers the ability to train the classifier with exemplary data that is more condensed and subject to less variability. The advantage of this is that better discrimination is possible between low levels of LOI and other categories. By recognizing the smaller variations in the distribution of MUP features for low LOI, sensitivity and specificity become better balanced. In contrast, when low LOI MUPs are not labeled as such, they become overshadowed by adjacent classes, resulting in a higher specificity at a cost of sensitivity.

7.1.2 Validation Using MDX Data

Similar observations were made in both the simulated data and real-world data. Specifically, the hierarchical method was able to significantly improve upon the balance of sensitivity and specificity. The magnitude of the improvement was slightly higher for the real-world data, but a large part of this

is attributed to having stratified the data into a smaller number of LOI categories, resulting in an easier classification problem.

Again, the LDA classifier with BR aggregation had the best overall performance. However, the contrast between this and using the PD classifier with AR aggregation was more obvious with this data set. In particular, when LDA was used, BR provided great performance but AR did not. Likewise when PD was used, AR provided good (but not great) performance while BR did not. These differences do not suggest that one method is better than the other, but rather highlight the different behaviors that depend on the quality of the evidence. When the standard disease stratification method was used, PD had slightly better performance than LDA in terms of better SSD. However, with the hierarchical method, only the AR method had significant improvement for PD because the quality of the evidence was less accurate. In other words, since there was fewer training data per class, a larger number of MUPs were incorrectly classified with false (inaccurate) confidence. The specific behavior of these aggregation methods will be discussed in greater detail in a later section of this chapter.

7.2 Accuracy of LOI Categorization

The optimal combinations of class structure and classification method are now discussed. For an LOI categorization method to be useful, it must be able to categorize LOI with acceptable accuracy, while reporting a sufficient range of clinically useful LOI states at the level of inference.

7.2.1 Evaluation Using Simulated Data

As seen in the results section, the accuracy of categorization when high-resolution stratification is used is quite poor. This is expected because, as seen in the distribution plots, there was a high degree of overlap in the feature distributions of the LOI groups. Thus, attempting to characterize clinical LOI at this level of stratification is not practical. The modified hierarchical strategy was able to improve on the accuracy of LOI detection, as long as the number of LOI groups was reduced from three to two. In particular, the strategy of re-labeling muscles categorized as having medium LOI proved to be most effective. The re-labeling was accomplished by selecting the second-highest characterization score and using that category as the assigned LOI label.

When the training data was re-labeled prior to training the second stage classifier, the performance improvement was not as significant. By doing so, the number of errors made between the CTRL group and the high LOI group increased, since the lower LOI groups were not represented as

concisely. Pattern Discovery did not provide as much of an increase in performance for this optimal LOI stratification. Again, this is likely due to an insufficient number of training samples when high-resolution stratification is used. Thus, it is more likely that PD would out-perform LDA if the training data for the low and medium groups was combined prior to classifier training.

The two-stage classifier method was used and is recommended because it allows for the possibility of different stratification strategies and aggregation techniques to support the two objectives of disease and LOI characterization. Further, by only using training data for the category predicted by the first stage, the number of classes in the second stage is reduced significantly, thus reducing computational complexity. In the case of two disease states and three LOI categories in each, the number of classes is reduced from seven to three or four, depending on the strategy.

This strategy represents a compromise between the number of clinical LOI states and the accuracy with which each of those states can be categorized.

7.2.2 Validation Using MDX Data

The MDX data could not be used to validate the hierarchical method when characterizing LOI. Since there were only two LOI categories present, it was not possible to reduce the stratification resolution and still report a clinical LOI state. Thus, only choices in stratification schemes were limited to whether or not the CTRL group was to be included. The two-stage classifier was shown to work best when only the two LOI states were used in classifier training. Thus the problem was reduced to a binary decision. The LDA classifier, using BR or AB aggregation methods, provided maximal performance. PD performance was slightly lower, but nonetheless acceptable. However, while the improvement for PD came at a cost of specificity, this was not the case with LDA. Thus, the ability for LDA to balance the accuracy in categorizing low and high LOI states while improving performance makes it a suitable candidate for the second stage LOI classifier. Because the classifier is making a binary decision, transparency of the method is less of an issue, and visualization of the decision space may even be possible with sufficient feature reduction.

7.3 Continuous Measures of LOI

Although previous work has shown strong correlation between muscle characterization confidence and LOI, the results presented here suggest that it is not a very useful measure. Even if a score is highly correlated, the overlap of scores across LOI groups, as depicted in the correlation plots, makes

it difficult to interpret an assigned score. At most, the range of score values might make it possible to establish certain intervals for a given LOI group, but the likelihood of a muscle falling into the correct interval would be quite low. As such, it is not clinically useful to report an LOI measure in this way.

The interpolated scores, while not being true conditional probabilities, did improve correlation. However, the degree of improvement still does not increase the utility of a continuous measure because they are still subject to the same amount of overlap. These limitations make it clear that continuous measures of LOI are not clinically viable.

With this in mind, a continuous measure is still sought. Specifically, when coupled with the discrete LOI classification studied in the previous section, a continuous measure could be thought of as a measure of confidence in the predicted LOI state. Thus, if two or more LOI states had high confidence, this might indicate that the true LOI state is somewhere in the middle of the range spanned by the specified LOI states. In contrast, if only a single LOI state had a high measure of confidence, then the true LOI would likely fall within this category. If such a reporting scheme were developed, then an interpolated continuous measure of LOI value could be incorporated into a user interface to aid in presenting the data.

The correlation results from the MDX data show much more promise as continuous measures, however these results are only limited to two LOI categories. Further study would be needed on larger clinical datasets in order to determine if these findings are significant.

7.4 Aggregation Measures

The behavior of aggregation measures as a function of the amount of evidence (number of MUPs) is discussed. By understanding the limitations of each method an appropriate method can be selected based on the level of conservativeness required by the system operator.

7.4.1 Evaluation Using Simulated Data

When using standard disease stratification, the ability for PD to discriminate between CTRL and NEUR-25 is slightly better than that of LDA. This is likely due to the fact that with this type of stratification, the number of training samples per class is adequate in order for PD to perform optimally. As such, PD is able to detect complex patterns in the training data that help discriminate between these categories. The linear decision boundary produced by LDA is not able to accomplish this as well. MUPs that fall into the low LOI region are assigned a lower conditional probability value

because they are closer to the discriminant function. As such, although a MUP might be representative of an exemplary MYO-25 MUP, the amount of evidence it provides to the decision is underweighted. As the amount of evidence is increased, there is a higher chance of low-confidence MUPs appearing due to a lack of a crisp definition of abnormality. Thus, there are fewer outliers present to influence the decision, and the AR method comes out on top. However, when 25 or more MUPs are considered, the chance of observing random outliers in either disease category is increased, and the likelihood of a miss-classification is increased, especially in the less conservative measures. Since PD is able to detect higher order patterns, it is capable of producing more accurate conditional probabilities. As such, the outlier MUPs allow the more extreme aggregation methods to converge on a decision. AR still performs reasonably, but under-weights these highly confident MUPs until enough of them are present (i.e., 25) in order to make a decision. The WMLO seems to provide a good compromise among the methods, performing reasonably well but not jumping to conclusions too early on.

The hierarchical method resulted in a significant improvement in accuracy for LDA. However, the AR method did not improve as much as the other aggregation measures. This indicates that LDA is producing better conditional probability estimates, and thus correct outliers, that allow the more extreme aggregators to make correct decisions with fewer pieces of evidence (i.e., smaller number of MUPs). The AR method requires many more MUPs (20 or more) before its accuracy is maximized, but then it drop at 25 MUPs. Given the mixture of MUPs from the normal and diseased categories present in a low LOI muscle, the AR method is too conservative and discounts the outliers that represent a diseased state.

For PD, the overall improvement seen by the hierarchical classifier is less pronounced. As well, the AR method performs better than the others which is the converse of the disease stratification method. By increasing the stratification resolution, PD suffers from a lack of training data and so the conditional probabilities are calculated less accurately. As such, the more extreme aggregators falsely rely on the presence of outliers. The improved performance of the AR method suggests that a more conservative aggregator is appropriate under these circumstances.

To summarize, BR and AB are quicker to jump to conclusions about the presence of abnormality, in a manner which is essentially akin to counting outliers. While this approach works well both in theory and practice, it is highly dependent on the assumption that the quality of the data (specifically

the estimates of conditional probability) is high. Averaging and methods based on it (i.e., WMLO) are more conservative and typically require more evidence before they make a decision. However, given the variability of MUPs within a muscle, the AR technique will always under-perform compared to the BR-based methods because it is less influenced by valuable outlier information, subject to the accuracy of such information. In contrast, if assurances about data quality cannot be made, BR-based methods will be more susceptible to errors and will not fail gracefully when a decision cannot be made with certainty. Under most circumstances, the aggregation methods seem to be ranked in the expected order, with the AB and WMLO methods between the AR and BR strategies. As such it might be reasonable to expect that either AB or WMLO would be a suitable choice for decision making when the methods are generalized to new data.

7.4.2 Validation Using MDX Data

The findings in the real-world dataset are closely aligned with those obtained from simulated data. A 'dip' in performance was also observed when using LDA with 25 MUPs. The explanation for this is the same as above. Similar trends were observed between the two classifiers, as well as across the different stratification methods. In particular, PD performed as well as LDA for higher MUP numbers, and the more extreme aggregators out-performed AR since confidence scores calculated by PD are more accurate when sufficient training data is provided. When the hierarchical method is used, a major improvement is seen for LDA for the extreme aggregators, just as in the simulated case. However, for PD, there is a significant drop in performance, although the AR method still does slightly better than the other methods. The drop in performance can also be explained by lower accuracy of conditional probability estimates, however it seems that there is another factor to consider: Looking at the distribution plots of this data set, there is considerable overlap between the CTRL and MYO-L groups. In the simulated case, the overlap is more symmetric across the LOI groups. However for the MDX data, there is more overlap between the CTRL and MYO-L groups than there is between the MYO-L and MYO-H groups. Thus there is greater confusion (and higher number of errors) made between the former pair. This, combined with the fact that the number of training samples are lower, results in less accurate estimations of MUP confidence (i.e., incorrect conditional probability estimates), as well as greater confusion in MUP characterizations. The combined effect results in lower performance that cannot be sufficiently compensated for by the conservative AR aggregation method.

Chapter 8
Conclusions

A hierarchical stratification scheme was evaluated and validated against real-world clinical data. When the method was applied to the characterization of clinical disease states, the hierarchical scheme was able to significantly improve upon standard classification methods. The improvement was realized as a balance in detection accuracy across muscle groups of various disease and LOI states, and not necessarily as an increase in total accuracy. In particular, the methods are capable of narrowing the trade-off between sensitivity and specificity. This performance improvement was noted in both the simulated and clinical data sets, despite the fact that they were stratified quite differently. The improvement is attributed to the fact that at a higher resolution of stratification, the subtle differences between LOI states are more easily detected. However, as was seen when using the PD classifier, greater stratification resolution increases the demand for training data, and further strata would likely lead to diminishing returns. Despite a reliance on training data, PD was able to perform quite well, and was found to be more robust when used in conjunction with the AR aggregation method.

A two-stage classifier was also developed so that an optimal configuration could be utilized for both disease and LOI characterization. The second stage consisted of a classifier that was re-trained using exemplary data specific to the predicted class from the first stage. In the case of simulated data, these reduced the number of classes involved in the classifier decision by almost a factor of two, resulting in much less complex class boundaries and reduced computational effort. When the hierarchical scheme was applied to LOI characterization, it was able to provide an effective compromise between categorization accuracy and class resolution. The optimal configuration was able to characterize LOI into two clinical states (low and high) at the level of inference with up to 77% accuracy for simulated data. The hierarchical scheme used for LOI categorization could not be validated on clinical data because such data was insufficiently stratified with respect to LOI. However, an optimal stratification scheme, consisting of a two-class LOI classifier was shown to categorize LOI with 89% accuracy.

Continuous measures of LOI were found to be impractical due to significant overlap in assigned muscles scores across different LOI categories. However, the interpolated scores produced by the hierarchical classifier might be useful for visualization purposes.

Aggregation measures were evaluated based on how they performed with varying amounts of evidence. In particular, as the number of MUPs was increased, the methods improved, with some exception. When the quality of the data was good, and confidence in MUP characterizations were high, the BR and AB methods were able to take advantage of the information contained in outliers in order to make an accurate decision with less evidence (i.e., fewer number of MUPs). However, when the quality of evidence is estimated less accurately, these methods were more likely to make an erroneous decision, while the conservative averaging technique was able to provide more consistent performance. A notable drawback was observed however, where all of the methods presented with a performance 'dip' as the amount of evidence increased to 25 MUPs. This 'dip' was only observed in cases where high levels of confusion were thought to exist within the MUP characterizations, rather than highly confident diagnostically relevant outliers. Further study is required to determine how additional evidence would affect the methods under these conditions. In general, however, none of the aggregation methods were found to be significantly better or worse than the others. In fact, they behave as expected and it is therefore difficult to draw a particular conclusion or recommendation. In order to balance the compromising effects of AR with the slightly higher accuracy of BR, it seems logical to select one of either WMLO or AB for implementation purposes. Since level of risk (reliance on information contained in outliers) allowed by AB can be varied by the choice the of λ parameter (see equation 3.3 and equation 3.4), it is the logical choice, giving the decision maker the ability to moderate the risk level as necessary.

In general, the ability to characterize neuromuscular disease was very promising, despite that fact that feature distributions plots indicate considerable overlap and non-Gaussian probability densities. These findings further support conclusions reached by several authors (Pino, 2009; Pino et al., 2010; Pino et al., 2008; Hamilton-Wright et al., 2010; Pfeiffer and Kunze, 1995) with regards to the advantages of probabilistic muscle characterization. In particular, the use of multivariate statistics and pattern recognition techniques is able to provide a robust framework to detect neuromuscular disorders. By aggregating evidence obtained from multiple statistical samplings of a muscle, an accurate inference can be made. Further, it is possible to transparently explain the findings in a manner that is consistent with a subjective approach. A significant focus of this work was to understand and describe the behavior of various aggregation methods that were analogous to certain human decision-making paradigms. Such understanding provides critical insight into the selection of such techniques when considering which method to include in a CDSS framework. Ultimately,

however, only general guidelines can be made at this point because the methods are greatly dependent on factors such as the quality, and availability of data. Further, an appropriate method should be selected to meet the specific goals of the decision: namely, a method being used to detect disease should use the appropriate level of conservatism depending on the costs associated with type I or type II errors. Likewise, when detecting LOI, it might be more reasonable to use a conservative approach rather than risking the chance of incorrectly identifying the severity of a disease.

The intention of this work was not to produce a specific recipe for success in disease characterization. Rather, the objectives were to understand the behaviors of the aggregation methods and the conditions for which each classifier and aggregation method was optimally suited. In doing so, it was found that LDA performed better than PD in terms of accuracy, but not by a large degree. This is likely due to the fact that the simplistic nature of LDA makes it better at estimating conditional probability with fewer data samples, but the validity of such a statement is likely highly dependent on the underlying distributions of the training data. In particular, LDA performs optimally when the classes obey a Gaussian distribution. The features presented in this thesis are clearly not Gaussian, but despite this LDA was able to perform quite well. On the other hand, PD was able to perform comparably under a variety of conditions. The extent to which PD's performance depends on the amount of training data available is not clear. Future work should assess the performance of PD as a function of the amount of training data, to determine if further improvements can be made, and whether or not there are further gains in using hierarchical methods.

As a general recommendation based on this work, the author suggests that PD is still the best classifier for use in a CDSS, mainly because of its transparency. However, LDA might be more appropriate in the second stage as an LOI classifier, because an explanation of these results is not as high of a priority. However, if transparency were desired, than attempts could be made to reduce the number of features used by LDA to perhaps two or three, allowing for easy visualization.

Future Work

In addition to the suggestions for future work already made, there are several questions that were not addressed by this work. The first of which has to do with the behaviors of aggregation measures. Aggregation measure behavior was only assessed in terms of accuracy thus far. While methods were considered for continuous LOI correlation, there was no emphasis on how the various methods estimate confidence. The ability to accurately report confidence is necessary for a CDSS, and the

paradigm used to arrive at a decision must be chosen based on the level of conservatism required. It would be useful to assess whether any particular aggregation method offers a more accurate estimate of conditional probability. Although there are methods to obtain well-calibrated confidence scores that represent true conditional probabilities (Pino, 2009), such calibration is cumbersome and undesirable in most circumstances. Thus it would be prudent to select an aggregation method that minimizes the need for calibration.

The methods centered on data stratification attempt to better represent variations due to different levels of LOI in the training set. However a radically different approach may be to attempt to purify the training data. Since it is well known that MUPs of all categories can be found within a particular set of data, one approach might be to cluster the training set in order to maximize the differences between classes. This would essentially 'purify' the feature distributions by increasing separation, allowing for higher estimates of confidence. Furthermore, in cases where further stratification of the data by an expert into LOI groups was unavailable, a clustering approach might be able to detect and label subtle changes that can be attributed to different levels of LOI. At this time, it is unclear how such techniques would be used, as clusters would be unlabeled and unordered, and future consideration is required.

Appendix A
Useful Derivations

A.1 The Inverse Logistic Function (*logit)*

The *logit* of a number p on the interval (0, 1) is given by:

$$logit(p) = \log(p) - \log(1-p) \qquad (A-1)$$

When p is a probability, then the term $p/(1 - p)$ is defined as the odds. Under this condition, the following results:

1. The *logit* of a probability is the logarithm of its odds:

$$logit(p) = log\left(\frac{p}{1-p}\right) \qquad (A-2)$$

2. The logarithm of an odds ratio (between two probabilities) can be expressed as the difference between their *logits*:

$$log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right) = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = logit(p_1) - logit(p_2) \qquad (A-3)$$

A.2 Bayes' Theorem Expressed in *logits*:

Bayes' theorem is given by:

$$P(y_k|MUP_1, MUP_2, \dots MUP_n) = \left(P(y_k)\prod_{i=1}^{N} P(MUP_i \mid y_k)\right) \times \left(\sum_{j=1}^{K}\left[\prod_{i=1}^{N} P(MUP_i \mid y_j)\right]P(y_j)\right)^{-1} \qquad (A-4)$$

and can be re-expressed in terms of log-odds as:

$$logit(y_k|MUP_1, MUP_2, \dots MUP_n) = logit(y_k) + \sum_{i=1}^{N} \frac{P(MUP_i \mid y_k)}{\sum_{j=1}^{K} P(MUP_i \mid y_j), j \neq k} \qquad (A-5)$$

When expressed in this way, it can be seen that Bayes' Theorem is additive with respect to each piece of evidence (MUP conditional probability), and the weight (significance) of each piece of evidence is proportional to its odds ratio.

Appendix B
Example of a CDSS

This appendix provides screenshots of a CDSS based on probabilistic muscle characterizations. The figures illustrate the advantages of the probabilistic muscle characterization frameworks proposed in (Pino, 2009), namely its ability to explain how each MUP template contributes to a muscle characterization and to explore the rationale for this contribution. The methods depicted herein allow a decision maker to 'drill down' from high levels of abstraction to detailed explanations.

Figures B-1 and B-2 provide visualizations related to MUP characterizations obtained using PD based estimates of conditional probability. Figure B-2 presents individual MUP characterizations as small pie charts, each of which is shown to contribute to the overall muscle characterization and level of involvement shown on the right. The muscle characterization scores and involvement percentages (using BR and AR, respectively) for each category are represented by large pie charts. Figure B-2 shows the feature events that underlie a single MUP characterization. The plots on the bottom left provide a visual interpretation of the rules generated by PD. Each shape represents a discrete feature value (low (L), medium (M) or high (H)), and shapes stacked along the same vertical line represent feature values that occurred together as an nth order pattern. Patterns for or against each category are displayed with their x-position reflecting the level of assertion or refutation provided by the respective rule. The solid red bars represent the combined assertion across all rules for a category. In this case, the evidence suggests that the muscle from which the MUPT was detected is likely myopathic, but also possesses normal and neurogenic characteristics. This confusion is driven by the presence of rules that provide evidence both for and against the myopathic and neurogenic categories. Generally, higher order patterns provide a higher weight of evidence and thus form the basis of the decision, in this case showing strong support for the myopathic category and strong refutation of the neurogenic category. A second order rule is also present in both the myopathic and neurogenic cases, providing contrary evidence to the higher order rule, but to a much lesser degree. The amount of evidence is less convincing for the normal category (shown by lower assertion values), but the presence of two separate rules bring the overall assertion further to the left. The overall characterization is still myopathic, but not with a very high level of confidence given the conflicting evidence. In Figure B-2, in the right column are plots of the distributions of the values for each feature. The three bins used for quantization (L, M, H) are demarcated, and green and red bars

represent the normal and abnormal ranges for each feature value. The white bar and plot represent the feature value and distribution of feature values for the MUP template and MUP template set, respectively. Figures B-1 and B-2 again demonstrate the ability of the probabilistic methods to provide a quantitative overall muscle characterization complete with a measure of confidence and the ability to see the distribution of MUP characterizations as well as the ability to 'drill down' into the details of the individual MUP characterizations. These examples demonstrate the advantages of using transparent decision support systems for QEMG.

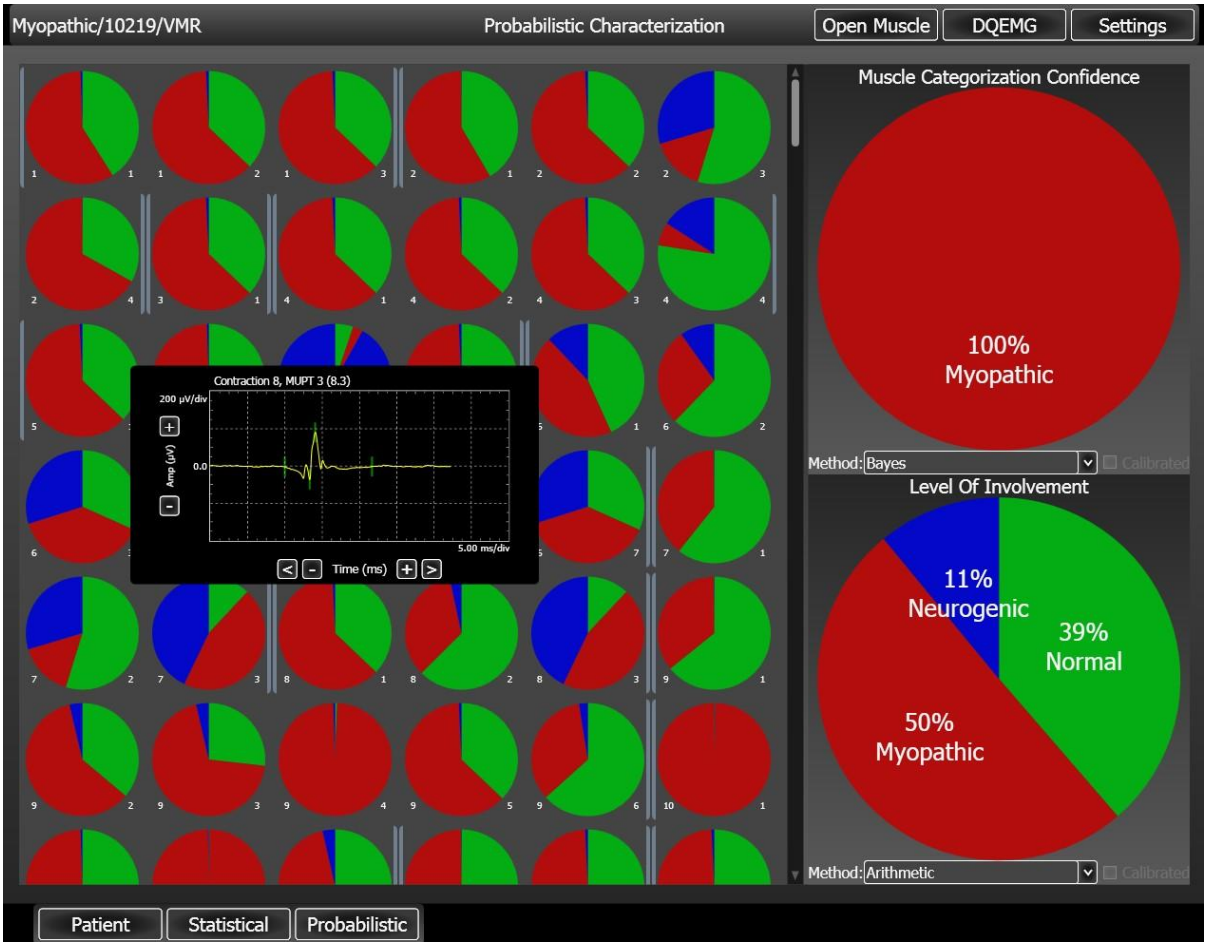

Figure B-1: Probabilistic Muscle Characterization: Small pie charts at left represent individual MUP characterizations. Large pie chart at top right represents muscle characterization and large pie chart at bottom right represents the estimated level of involvement.

Figure B-2: Probabilistic Muscle Characterization: Details of a specific MUP characterization with support for the myopathic category.

References

Buchthal, F., Pinell P., and Rosenfalck, P. 1954. Action potential parameters in normal human muscle and their physiological determinants. *Acta Physiologica Scandinavica* 32, no. 2 (November): 219-229.

Budescu, David V., and Hsiu-Ting Yu. 2006. To Bayes or Not to Bayes? A Comparison of Two Classes of Models of Information Aggregation. *Decision Analysis* 3, no. 3: 145-162.

Christodoulou, C.I., and Pattichis, C.S.. 1999. Unsupervised pattern recognition for the classification of EMG signals. *IEEE Transactions on Biomedical Engineering* 46, no. 2 (February): 169-178. doi:10.1109/10.740879.

Clemen, Robert T., and Reilly T. 2000. *Making Hard Decisions with DecisionTools*. 1st ed. Duxbury/Thomson Learning, June 23.

Daube, Jasper R, and Rubin, Devon I. 2009. Needle electromyography. *Muscle & Nerve* 39, no. 2 (February): 244-270. doi:10.1002/mus.21180.

Doherty, T.J., and Stashuk D.W. 2003. Decomposition-based quantitative electromyography: Methods and initial normative data in five muscles. *Muscle & Nerve* 28, no. 2: 204–211. doi:10.1002/mus.10427.

Duda, R.O., Peter E. Hart, and David G. Stork. 2000. *Pattern Classification*. 2nd ed. Wiley-Interscience, October 26.

Economou, G.P., D Lymberopoulos, E Karavatselou, and C Chassomeris. 2001. A new concept toward computer-aided medical diagnosis--a prototype implementation addressing pulmonary diseases. *IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society* 5, no. 1 (March): 55-66.

Economou, G.-P.K., P.D. Goumas, and K. Spiropoulos. 1996. A novel medical decision support system. *Computing & Control Engineering Journal* 7, no. 4: 177-183. doi:10.1049/cce:19960404.

Fuglsang-Frederiksen, A., B. Johnsen, S. Vingtoft, and J. Rønager. 1993. An EMG expert system — KANDID and a new EMG database structure. *Electroencephalography and Clinical Neurophysiology* 87, no. 2: 17-.

Gevaert, O., F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor. 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22, no. 2006: 184-.

Hamilton-Wright, A., and D. W Stashuk. 2006. Clinical characterization of electromyographic data using computational tools. *(2006). 2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (IEEE Cat. No.06EX1522C)(pp.7 pp.). Piscataway*: 7 pp.

Hamilton-Wright, Andrew, Linda McLean, Daniel W Stashuk, and Kristina M Calder. 2010. Bayesian aggregation versus majority vote in the characterization of non-specific arm pain based on quantitative needle electromyography. *Journal of NeuroEngineering and Rehabilitation* 7, no. 13: 12 pp.

Katsis, Christos D., Themis P. Exarchos, Costas Papaloukas, Yorgos Goletsis, Dimitrios I. Fotiadis, and Ioannis Sarmas. 2007. A two-stage method for MUAP classification based on EMG decomposition. *Comput. Biol. Med.* 37, no. 9: 1232–1240.

Kononenko, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* 23, no. 1 (August): 89-109.

McGill, K C, K Lau, and L J Dorfman. 1991. A comparison of turns analysis and motor unit analysis

in electromyography. *Electroencephalography and Clinical Neurophysiology* 81, no. 1 (February): 8-17.

Pfeiffer, G. 1999. The diagnostic power of motor unit potential analysis: an objective bayesian approach. *Muscle & Nerve* 22, no. 5 (May): 584-591.

Pfeiffer, G., and K. Kunze. 1995. Discriminant classification of motor unit potentials (MUPs) successfully separates neurogenic and myopathic conditions. A comparion of multi- and univariate diagnostical algorithms for MUP analysis. *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control* 97, no. 5 (October): 191-207. doi:10.1016/0924-980X(95)00072-0.

Pino, L. 2009. Neuromuscular clinical decision support using motor unit potentials characterized by 'pattern discovery'. Canada: University of Waterloo (Canada).

Pino, L.J., D.W. Stashuk, S.G. Boe, and T.J. Doherty. 2008. Motor unit potential characterization using "pattern discovery". *Medical Engineering & Physics* 30, no. 5 (June): 563–573. doi:10.1016/j.medengphy.2007.06.005.

———. 2009. Chapter 25 Decision support for QEMG. In *Motor Unit Number Estimation (MUNE) and Quantitative EMG: Selected Presentations from the Second International Symposium on MUNE and QEMG, Snowbird, Utah, USA, 18-20 August 2006*, Volume 60:247-261. Elsevier.

———. 2010. Probabilistic muscle characterization using QEMG: Application to neuropathic muscle. *Muscle & Nerve* 41, no. 1: 18-31. doi:10.1002/mus.21456.

Pino, L.J., D.W. Stashuk, and S. Podnar. 2008. Bayesian characterization of external anal sphincter muscles using quantitative electromyography. *Clinical Neurophysiology* 119, no. 10 (October): 2266-2273. doi:10.1016/j.clinph.2008.06.017.

Pino, L.J., D.W. Stashuk. 2008. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. In , 4138-4141. Vancouver, BC, 8. doi:10.1109/IEMBS.2008.4650120.

Podnar, S, and D B Vodusek. 2001. Standardization of anal sphincter electromyography: utility of motor unit potential parameters. *Muscle & Nerve* 24, no. 7 (July): 946-951.

Podnar, Simon. 2004a. Usefulness of an increase in size of motor unit potential sample. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 115, no. 7 (July): 1683-1688. doi:10.1016/j.clinph.2004.02.016.

———. 2004b. Criteria for neuropathic abnormality in quantitative anal sphincter electromyography. *Muscle & Nerve* 30, no. 5 (November): 596-601. doi:10.1002/mus.20148.

———. 2005. Comparison of different outlier criteria in quantitative anal sphincter electromyography. *Clinical Neurophysiology* 116, no. 8 (August): 1840-1845. doi:10.1016/j.clinph.2005.04.023.

———. 2008. Comparison of parametric and nonparametric reference data in motor unit potential analysis. *Muscle & Nerve* 38, no. 5 (November): 1412-1419. doi:10.1002/mus.21102.

———. 2009a. Predictive values of the anal sphincter electromyography. *Neurourology and Urodynamics* 28, no. 8: 1034-1035. doi:10.1002/nau.20728.

———. 2009b. Predictive values of motor unit potential analysis in limb muscles☆. *Clinical Neurophysiology* 120, no. 5 (5): 937-940. doi:10.1016/j.clinph.2009.02.165.

Podnar, Simon, and Mićo Mrkaić. 2002. Predictive power of motor unit potential parameters in anal sphincter electromyography. *Muscle & Nerve* 26, no. 3: 389-394. doi:10.1002/mus.10207.

Podnar, Simon, and Mićo Mrkaić. 2003. Size of motor unit potential sample. *Muscle & Nerve* 27, no. 2 (February): 196-201. doi:10.1002/mus.10310.

Preston, D.C, and B.E Shapiro. 2002. Needle electromyography Fundamentals, normal and abnormal patterns 20: 361-396.

Revett, K., F. Gorunescu, M. Gorunescu, E. El-Darzi, and M. Ene. 2005. A breast cancer diagnosis system: a combined approach using rough sets and probabilistic neural networks. *(2005). EUROCON 2005-The International Conference on 'Computer as a Tool'(pp.4 pp.). Piscataway*: 4 pp.

Salzberg, Steven L. 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning* 16, no. 3 (9): 235-240. doi:10.1007/BF00993309.

Schizas, C N, L T Middleton, and C S Pattichis. n.d. Neural network models in EMG diagnosis. *IEEE Transactions on Biomedical Engineering* 42, no. 5: 486-496.

Sonoo, Masahiro. 2002. New attempts to quantify concentric needle electromyography. *Muscle & Nerve* Suppl 11: S98-S102. doi:10.1002/mus.10154.

Šprogar, Matej, Mitja Lenič, and Silvia Alayon. 2002. Evolution in Medical Decision Making. *J. Med. Syst.* 26, no. 5: 479-489.

Stålberg, E, C Bischoff, and B Falck. 1994. Outliers, a way to detect abnormality in quantitative EMG. *Muscle & Nerve* 17, no. 4 (April): 392-399. doi:10.1002/mus.880170406.

Stålberg, E, S Stålberg, M Melander, and K Arimura. 1991. A personal computer based system used in electromyography for interpretation and reporting. *Computer Methods and Programs in Biomedicine* 34, no. 2 (March): 219-227.

Stashuk, D.W.. 2001. EMG signal decomposition: how can it be accomplished and used? *Journal of Electromyography and Kinesiology: Official Journal of the International Society of Electrophysiological Kinesiology* 11, no. 3 (June): 151-173.

———. 1999. Decomposition and quantitative analysis of clinical electromyographic signals. *Medical Engineering & Physics* 21, no. 6 (September): 389-404.

———. 1993. Simulation of electromyographic signals. *Journal of Electromyography and Kinesiology* 3, no. 3 (September): 157–173. doi:10.1016/S1050-6411(05)80003-3.

Stashuk, D.W., and W.F Brown. 2002. Quantitative Electromyography. In *Neuromuscula Function and Disease*.

Stewart, C R, S D Nandedkar, J M Massey, J M Gilchrist, P E Barkhaus, and D B Sanders. 1989. Evaluation of an automatic method of measuring features of motor unit action potentials. *Muscle & Nerve* 12, no. 2 (February): 141-148.

Suojanen, M, S Andreassen, and K G Olesen. 2001. A method for diagnosing multiple diseases in MUNIN. *IEEE Transactions on Bio-Medical Engineering* 48, no. 5 (May): 522-532. doi:10.1109/10.918591.

Whitlock, M C. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* 18, no. 5 (September): 1368-1373. doi:10.1111/j.1420-9101.2005.00917.x.

Wong, Andrew K. C., and Yang Wang. 1997. High-Order Pattern Discovery from Discrete-Valued Data. *IEEE Trans. on Knowl. and Data Eng.* 9, no. 6: 877-893.

Xiang, Zuoshuang, Rebecca M. Minter, Xiaoming Bi, Peter J. Woolf, and Yongqun He. 2007. miniTUBA: medical inference by network integration of temporal data using Bayesian analysis. *Bioinformatics* 23, no. 18 (September 15): 2423-2432. doi:10.1093/bioinformatics/btm372.

Xie, Hongbo, Hai Huang, and Zhizhong Wang. 2005. Multiple Feature Domains Information Fusion for Computer-Aided Clinical Electromyography. In *Computer Analysis of Images and Patterns*, 304-312. http://dx.doi.org/10.1007/11556121_38.

Zalewska, E., I. Hausmanowa-Petrusewicz, and E. Stålberg. 2004. Modeling studies on irregular motor unit potentials. *Clinical Neurophysiology* 115, no. 3 (March): 543-556.

doi:10.1016/j.clinph.2003.10.031.
Zalewska, Ewa, and Irena Hausmanowa-Petrusewicz. 2005. The SIIR index--a non-linear combination of waveform size and irregularity parameters for classification of motor unit potentials. *Clinical Neurophysiology* 116, no. 4 (April): 957-964. doi:10.1016/j.clinph.2004.11.012.