A Multiple Instance Learning Approach to Electrophysiological Muscle Classification for Diagnosing Neuromuscular Disorders Using Quantitative EMG

by

Tahereh Kamali

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

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dawn MacIsaac Associate Professor, Dept. of Computer Science/Engineering, University of New Brunswick
Supervisor(s):	Daniel W. Stashuk Professor, Dept. of Systems Design Engineering, University of Waterloo
Internal Member:	David Clausi Professor, Dept. of Systems Design Engineering, University of Waterloo
Internal-External Member:	Mark Crowley Assistant Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Other Member(s):	Ning Jiang Assistant Professor, Dept. of Systems Design Engineering, University of Waterloo

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 disorder is a broad term that refers to diseases that impair muscle functionality either by affecting any part of the nerve or muscle. Electrodiagnosis of most neuromuscular disorders is based on the electrophysiological classification of involved muscles which in turn, is performed by inferring the structure and function of the muscles by analyzing electromyographic (EMG) signals recorded during low to moderate levels of contraction. The functional unit of muscle contraction is called a motor unit (MU). The morphology and physiology of the MUs of an examined muscle are inferred by extracting motor unit potentials (MUPs) from the EMG signals detected from the muscle. As such, electrophysiological muscle classification is performed by first characterizing extracted MUPs and then aggregating these characterizations.

The task of classifying muscles can be represented as an instance of a multiple instance learning (MIL) problem. In the MIL paradigm a bag of instances shares a label and the instance labels are hidden, contrary to standard supervised learning, where each training instance is labeled. In MIL-based muscle classification, the instances are the MUPs extracted from the EMG signals of the analyzed muscle and the bag is the muscle. Detecting and counting the MUPs indicating a specific category of neuromuscular disorder can result in accurately classifying the examined muscle. As such, three major issues usually arise: how to infer MUP labels without full supervision; how the cardinality relationships between MUP labels contribute to predict the muscle label; and how the muscle as a whole entity is classified. In this thesis, these three challenges are addressed.

To this end, an MIL-based muscle classification system is proposed that has five major steps: 1) MUPs are represented using morphological, stability, and novel near fiber parameters as well as spectral features extracted from wavelet coefficients. This representation helps to analyze MUPs from a variety of aspects. 2) MUP feature selection using unsupervised similarity preserving Laplacian score which is independent of any learning algorithm. Hence, the features selected in this work can be used in other electrophysiological muscle classification systems. 3) MUP clustering using a novel clustering algorithm

called Neighbourhood Distance Entropy Consistency (NDEC) which contributes to solve the traditional problem of finding representations of MUP normality and abnormality and provides a dynamic number of MUP characterization classes which will be used instead of the conventional three classes (i.e. normal, myopathic, and neurogenic). This clustering was performed to highlight the effects of disease on both fiber spatial distributions and fiber diameter distributions, which lead to a continuity of MUP characteristics. These clusters can potentially represent several concepts of MUP normality and abnormality. 4) Muscle representation by embedding its MUP cluster associations in a feature vector, and 5) Muscle classification using support vector machines or random forests.

Quantitative results obtained by applying the proposed method to four electrophysiologically different groups of muscles including proximal arm, proximal leg, distal arm, and distal leg show the superior and stable performance of the proposed muscle classification system compared to previous works. Additionally, modelling electrophysiological muscle classification as an instance of the MIL can solve the traditional problem of characterizing MUPs without full supervision. The proposed clustering algorithm in this work, can be used as an effective technique in other pattern recognition and medical diagnostic systems in which discovering natural clusters within data is a necessity.

Keywords: Electrophysiological Muscle Classification, Multiple-Instance Learning, Needle Electrodiagnostic Examination, Neighbourhood Distance Entropy Consistency

Acknowledgements

I would like to thank my supervisor, Prof. Daniel W. Stashuk, for his guidance, support, insightful discussions and constructive feedback throughout my PhD studies. He has been a great mentor. I would like to thank him for giving me the opportunity to explore my research interests and helping me grow as an independent researcher.

I am deeply thankful to all my thesis committee members, Prof. David Clausi, Prof. Mark Crowley, and Prof. Ning Jiang. It has been a pleasure to have them on my committee. I am thankful to Prof. Ali Ghodsi and Prof. Ming Li who taught advanced graduate level courses on machine learning. The discussions with them were also very helpful and helped me in shaping my research ideas.

I have had a chance to meet many wonderful friends and enjoy their friendship since the beginning of my stay in Waterloo. I would like to thank all of them with whom I enjoyed many conversations on grad life and research.

Most importantly, I would like to express my deepest gratitude and love to my family for their unconditional love and support. My special gratitude and love goes to my dearest parents, Marzieh and Karim, my beloved husband Ameen, my sweetest flower Tasnim, my sister Razieh, and my brother Mohammad for all their continuous support and encouragement.

And, before and after everything, and in the midst of everything, All praise goes to God, the Lord of the worlds.

Dedication

To my family: without your love, patience, and help, this thesis wouldn't exist.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN Artificial Neural Network
ANOVA Analysis of Variance
ARI Adjusted Rand Index
ASAD Arbitrary Shape and Arbitrary Density
ASSD Arbitrary Shape and Specific Density

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies

CLARANS A Method for Clustering Objects for Spatial Data Mining
CURE Clustering Using Representatives

DBCV Density-Based Clustering Validation
DBSCAN Density-Based Spatial Clustering of Applications with Noise
DenClue Fast Clustering based on Kernel Density Estimation
DLT Deltoid
DPC Density Peaks Clustering
DT Decision Tree
DTI Diffusion Tensor Imaging
DWT Discrete Wavelet Transform

EAR Event Association Rules
EMC Electrophysiological Muscle Classification
EMG Electromyographic

FDI First Dorsal Interosseous
Fmajor Forceps Major

GA Genetic Algorithm
GDD Grid-based Clustering Algorithm for Multi-Density
GDI Global Density Information
GEI Global Entropy Information
GMMs Gaussian Mixture Models

IFO Inferior Fronto-Occipital Fasciculus
ILF Inferior Longitudinal Fasciculus

KNN k-Nearest Neighbours

LDA Linear Discriminant Analysis
LDI Local Density Information
LS Laplacian Score

MFP Muscle Fiber Potential
MIL Multiple Instance Learning
MST Minimum Spanning Tree
MU Motor Unit
MUP Motor Unit Potential
MUPT Motor Unit Potential Train
Myo Myopathy

NDEC Neighbourhood Distance Entropy Consistency
Neuro Neuropathy
NF Near Fiber
NMI Normalized Mutual Information
NN Nearest Neighbour
NSE Normalized Sub-band Energy

OPTICS Ordering Points To Identify the Clustering Structure
ORL Olivetti Face Dataset

PAM Partitioning Around Medoids
PD Pattern Discovery

QEMG Quantitative EMG Analysis

RF Random Forest

SC Spectral Clustering
Spc Specificity
SSN Shared Nearest Neighbours
SSSD Specific Shape and Specific Density
SVM Support Vector Machine

TA Tibialis Anterior

UEF University of Eastern Finland

VM Vastus Medialis

WaveCluster Wavelet-based Clustering
WM White Matter

Chapter 1

Introduction and Motivation

1.1 Summary

During the last thirty years, Electromyographic (EMG) signals have been widely utilized by clinicians and researchers as a versatile tool for the accurate and timely diagnosis of a variety of neuromuscular disorders. Historically EMG signals have been analyzed qualitatively; however, over the last decade, a great deal of interest has been found in quantitative EMG analysis (QEMG), in which a set of quantitative features of an EMG signal are extracted and assessed for their diagnostic information. This transition can be attributed to the fact that qualitative assessment is subjective and depends on clinician skill and experience. Furthermore, longitudinal studies which focus on estimating disease severity and progression are impossible using qualitative assessment, whereas, quantitative analysis is objective and provides tools to facilitate the completion of longitudinal studies.

The main purpose of this thesis is to propose methods which can improve the clinical utility of quantitative electromyographic techniques. For this purpose, this thesis focuses on proposing new methods to boost the performance of quantitative electrophysiological muscle classification which in turn can lead to viable diagnosis of neuromuscular disorders.

To date, several electrophysiological muscle classification systems have been developed using a supervised learning approach. Each of them has its own strengths and weaknesses due to the complexity of the problem. Generally speaking, standard supervised learning has many limitations with regard to electrophysiological muscle classification. In particular, providing labels for all examples in electrophysiological muscle training data, is not feasible. Instead, labels can be provided for groups of examples. In supervised learning terminology, the examples are called instances and the groups are called bags. The learning scenario where only labeled bags are available is called multiple instance learning (MIL). In this thesis, electrophysiological muscle classification is formulated as an instance of a MIL problem in which labels are only provided at the bag level in the electrophysiological muscle training set.

1.2 Overview of Electrophysiological Background

Electrophysiological muscle classification is a crucial step in the diagnosis of neuromuscular disorders and can be performed to assist discrimination between healthy muscles and those which are affected by a neuromuscular disease process. Physicians classify muscles based on the results of different clinical examinations and tests. Among them, EMG examinations, which study electrical potentials detected in a muscle at rest or during activation, are the most widely accepted [5]. Physiological and morphological aspects of a muscle are represented in the EMG signals recorded from that muscle [6]. As a result, analyzing the EMG signals assists in discovering possible alterations in both the physiology and structure of the underlying muscle [2].

Skeletal muscle is composed of numerous multinucleated densely packed muscle fibers. Each motor neuron innervates a set of different muscle fibers. The muscle fibers innervated by one motor neuron plus the neuron cell body, the long axon running down a motor nerve and its terminal branches, together constitute a motor unit (MU). In a healthy MU, discharge of the motor neuron subsequently leads to concurrent and consistent activation

of all MU fibers. The currents associated with the propagating muscle fiber action potentials flow throughout the extracellular space. Electrodes inserted inside the territory of a discharging MU detect potentials generated by these currents [5].

Active single muscle fibers generate waveforms called muscle fiber potentials (MFPs) and the summation of potentials generated by the different fibers of the same motor unit are called motor unit potentials (MUPs). To maintain or increase the force produced by a muscle, MUs are activated repeatedly. As such, each MU generates a train of MUPs which is called a motor unit potential train (MUPT). An EMG signal is comprised of the summation of the MUPTs detected during a contraction [7].

Neuromuscular disorders can be categorized into two main groups including myopathic and neurogenic. Myopathy is a group of disorders that is caused by the death or atrophy of muscle fibers, whereas the neuropathy refers to any disorders that is caused by death of or damage to the motor neurons. The primary symptom in myopathies is muscle weakness which is the result of dysfunctional and/or lost muscle fibers. In neuropathies, loss of motor neurons is an early sign, where fibers associated with degenerating motor neurons, lose their neuronal connection and become denervated. Hence, surviving healthy motor neurons in the immediate vicinity of the orphaned/denervated muscle fibers grow new axonal sprouts and re-innervate the denervated muscle fibers [8].

Physicians diagnose neuromuscular disorders by considering characteristics of a set of examined muscles. An individual muscle is electrophysiologically characterized, as normal, myopathic or neurogenic, by qualitatively analyzing sets of MUPs representing MUs sampled in the examined muscle and then aggregating their MUP characterizations. Qualitative muscle classification is subjective and depends on clinician skill and experience. Furthermore, longitudinal studies, which focus on estimating disease severity and progression, are difficult using such qualitative assessment.

In contrast, quantitative electrophysiological muscle classification is objective and provides tools to facilitate the completion of longitudinal studies [5]. Quantitative electrophysiological muscle classification systems, like physicians, consider sets of MUPs representing

sampled MUs, and quantitatively characterize individual MUPs and then aggregate these characterizations to obtain a muscle classification.

1.3 Challenges and Objectives of this Thesis

Designing a robust, and accurate quantitative electrophysiological muscle classifier is not straightforward due to the complex nature of the problem. Despite previous attempts, several open challenges still exist. The main challenges of this task are briefly explained below:

1. Non-Uniform Level of Disease Involvement: The level of disease involvement varies across constituent motor units of a muscle. As an example, a myopathic muscle can have severely affected, slightly affected, and normal motor units. Hence, an individual MUP cannot provide sufficient diagnostic information and electrophysiological muscle classification must be based on a set of MUPs produced by a representative sample of a muscle's MUs.

2. Partially Labelled Training Data: Most electrophysiological muscle classification training data is only partially labelled. Associated muscle labels are provided; however, individual MUPs do not have any labels. In some cases MUPs are labelled by physicians, but this is a time consuming and expensive task that is prone to errors due to the high volume of data.

3. Irrelevant Instances: There might be some MUPs recorded from a muscle that do not convey any information about the class label of the muscle, or these MUPs may be even more related to other classes of muscles. As an example, myopathic and neurogenic muscles will most likely generate several normal MUPs.

4. Dynamic Number of MUP Characterization Classes: Existing muscle classification methods consider only three classes (i.e., normal, myopathic and neurogenic) for MUP characterization as well as muscle classification. It is worth noting that fiber spatial

distributions and fiber diameter distributions in a MU are all modified by disease in a continuous way. Therefore, changes induced by the disease process is also continuous in MUP characteristics. As a result, except for extreme cases, distinct boundaries between MUPs produced by normal, myopathic and neurogenic MUs cannot be identified. Consequently, considering more than three classes for MUP characterization may be better able to reflect the various effects of disease.

5. Multi-Class Imbalanced Data: The electrophysiological muscle classification data distribution is highly skewed since representatives of the normal class appear much more frequently. On the other hand, the minority classes (i.e. myopathy and neuropathy) are more important from a diagnostic perspective.

6. High Leverage Apparent Outlier Observations: Due to the nature of the disease processes, there are some minority samples that represent clear cases of myopathic or neurogenic MUs and despite their rareness, they carry significantly important and useful diagnostic information. Hence, the distinction between true outliers and apparent outliers is both challenging and important.

The main purpose of this research is to propose methods which address the above challenges and boost quantitative electrophysiological muscle classification performance which in turn can lead to viable diagnosis of neuromuscular disorders. To this end, it is difficult to formulate the muscle classification problem in a standard supervised learning setting in which both training and test data are represented as individual feature vectors. Instead, the training and test data can be represented by sets or bags of feature vectors or instances. Hence, the task of classifying muscles is represented as an instance of a multiple instance learning (MIL) problem. In the MIL paradigm a bag of instances has a label and the instance labels are hidden, contrary to standard supervised learning, where each training instance is labelled. In MIL-based muscle classification, the instances are the MUPs extracted from the EMG signals of the analyzed muscle and the bag is the muscle.

1.4 Thesis Contribution

This dissertation contributes to electrophysiological muscle classification (EMC) by proposing novel methods to solve the traditional problem of MUP characterization without full supervision. The detailed contributions of this work are highlighted as follows.

- **A novel MIL framework to model electrophysiological muscle classification:** A machine learning literature survey [9] shows that MIL encompasses two different learning scenarios: MIL with the purpose of labeling bags, and MIL with the purpose of labeling instances. Classifiers that optimize performance on bags, may not provide the best possible instance labels. Additionally, these instance labels may change and be unstable under different training phases. These behaviours are not desirable, especially in cases where the instance labels carry medical significance. Hence, a potential MIL-based EMC system should by design provide the stability of instance labels as well as a high level of accuracy. We propose an MIL framework which provides stable instance (MUP) labels and accurate and robust bag (muscle) classification results.

- **A novel density-based clustering algorithm to discover natural clusters:** Traditional clustering algorithms model the clustering problem as an optimization task, in which the objective is defined based on minimizing specific metrics. These algorithms are limited to find clusters with convex polytopes. In contrast, density-based clustering algorithms aim at overcoming this limitation and try to partition data objects into meaningful groups that have relatively high density separated by low-density regions. We propose a new clustering algorithm that improves upon previous density-based clustering approaches. To this end, a dynamic density-based clustering algorithm called Neighbourhood Distance Entropy Consistency (NDEC) is proposed which simultaneously uses local and global density consistency information as well as consistency of neighbourhood distance entropy to discover the intrinsic clustering structure with arbitrary shape, size, and density.

NDEC is capable of identifying outliers and does not require prior knowledge about the number of clusters. In addition, it is not sensitive to initialization since the starting conditions are the same for all runs of the algorithm. Experiments on synthetic and real benchmark clustering datasets have demonstrated the efficiency and effectiveness of the NDEC method. Comparisons with k-means, DBSCAN, OPTICS, and Density Peaks clustering algorithms further show that NDEC can successfully discover natural clusters. Additionally, in this thesis, the utility of NDEC is demonstrated with its application on two real-world problems including segmentation of white matter tracts in diffusion tensor imaging and characterizing motor unit potential trains extracted from electromyographic signals.

- **A novel method to infer MUP labels without full supervision:** Electrophysiological muscle classification involves characterization of extracted motor unit potentials (MUPs) followed by the aggregation of these MUP characterizations. Existing techniques consider three classes for both MUP characterization and electrophysiological muscle classification. However, disease-induced MUP changes are continuous in nature, which makes it difficult to find distinct boundaries between normal, myopathic and neurogenic MUPs. Hence, MUP characterization based on more than three classes is better able to represent the various effects of disease.

  In this thesis, an MIL-based electrophysiological muscle classification system is presented which considers a dynamic number of classes for characterizing MUPs. To this end, NDEC is utilized to find clusters with arbitrary shape and density in a MUP feature space. These clusters represent several concepts of MUP normality and abnormality and are used for MUP characterization instead of the conventional three classes (i.e., normal, myopathic, and neurogenic).

- **A novel method for muscle classification which relies on the characterization of MUPs:** The electrophysiological muscle classification problem is naturally formulated using the MIL setting and needs an adaptation of standard supervised classifiers for the purpose of training and evaluating on the bags of instances. We

propose a novel MIL-based EMC system in which the muscle classifier uses the predictions made on MUPs to infer muscle labels. Quantitative results show the superior and stable performance of the proposed MIL-based EMC system compared to previous works performed with other supervised, semi-supervised and unsupervised methods.

There are three journal publications and one under review related to the main contributions of this work:

1. T. Kamali and D. W. Stashuk, Electrophysiological Muscle Classification using Multiple Instance Learning and Unsupervised Time and Spectral Domain Analysis, IEEE Transactions on Biomedical Engineering, 2018.

2. T. Kamali and D. W. Stashuk, A Density-Based Clustering Approach to Motor Unit Potential Characterizations to Support Diagnosis of Neuromuscular Disorders, IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, (7), pp. 956-966, 2017.

3. T. Kamali and D. W. Stashuk, Automated Segmentation of White Matter Fiber Bundles using Diffusion Tensor Imaging Data and a New Density based Clustering Algorithm, Artificial Intelligence in Medicine 73, pp.14-22, 2016.

4. T. Kamali and D. W. Stashuk, NDEC: A Density-Based Clustering Algorithm using Neighbourhood Distance Entropy Consistency, IEEE Transactions on Knowledge and Data Engineering, 2018 (Under Review).

1.5 Thesis Organization

There are seven chapters in this thesis including the Introduction. Chapter 2 provides a brief overview of muscle anatomy and physiology, neuromuscular disorders and electro-

physiological basics of EMG signal generation. In addition, previous works done to classify muscles using quantitative electromyography are summarized and a brief review of previous MIL techniques is presented. Chapter 3 states and formulates the electrophysiological muscle classification problem and provides an overview of the framework proposed in this work. Chapter 4 presents the basic concepts of cluster analysis and reviews briefly classical clustering algorithms from the literature. In addition, Neighbourhood Distance Entropy Consistency (NDEC) clustering algorithm is introduced and evaluated using a variety of artificial and real benchmark clustering datasets. In Chapter 5, the MIL-based EMC system using supervised time domain analysis is presented and evaluated on 103 sets of MUPs recorded in tibialis anterior muscles. In Chapter 6, the MIL-based EMC system using unsupervised time and spectral domain analysis is presented and assessed on 63, 83, 93, and 84 sets of MUPs recorded in deltoid, vastus medialis, first dorsal interosseous, and tibialis anterior muscles, respectively. Chapter 7 concludes the thesis and discusses opportunities for future work.

Chapter 2

Background and Related Work

This chapter presents an overview of the basic physiology of muscle contraction and EMG signal generation and detection techniques used for clinical applications as well as a brief description of EMG decomposition. Understanding these fundamental electrophysiological concepts assists in the appreciation of and provides a better insight into the methods presented in the next chapters. Furthermore, in this chapter, a review of existing quantitative EMC techniques is presented and the weaknesses and strengths of each technique are discussed. Finally, a brief review of current multiple instance learning classification techniques is presented.

2.1 Background

2.1.1 EMG Generation and Detection

The skeletal muscle is composed of numerous multinucleated densely packed muscle fibers that are surrounded by a thin layer of connective tissue. A muscle fiber is a very thin thread which has a length ranging from a few millimeters to 30 cm and a diameter of 10

μm to 100 μm. Each muscle fiber is connected to a motor neuron via the synapses of its neuromuscular junction [3].

The motor neuron cell body is located in the ventral horn of the spinal cord. Each motor neuron innervates a set of different muscle fibers. The muscle fibers innervated by one motor neuron plus the neuron cell body, the long axon running down a motor nerve and its terminal branches, together constitute a motor unit (MU). The muscle fibers of a healthy MU are spatially randomly distributed within a portion of the cross-sectional area of a muscle (i.e. the motor unit territory) [10].

When a MU (demonstrated in Fig. 2.1) is activated, all fibers in the MU contract and produce force. Groups of MUs work together to coordinate the contractions of a single muscle. Muscles that require precision and fine movement control usually have many MUs with a small number of fibers in each unit. There is an orderly recruitment of MUs. Small, slowly contracting, fatigue resistant MUs, are first recruited and produce small forces. With increasing force demands, large, fast contracting fatigable MUs join in [4].

Slower MUs are thereby more frequently used than faster ones. Muscle fibers possess the property of being excitable. An action potential is an electrical impulse transmitted from a nerve fiber branch to a muscle fiber at its neuromuscular junction. The electrical impulse is transferred by release of a specific type of neurotransmitter molecules which diffuse from the synapses of the motor nerve to the receptors on the plasma membrane of the muscle fiber and cause the membrane to generate its own action potential which travels along the muscle fiber at a rate of about 2 to 5 meters per second [5], [11].

In healthy muscles, an active motor neuron concurrently stimulates all muscle fibers connected to it. The currents corresponding to these potentials propagate throughout the extracellular muscle volume. Note that muscle fiber conduction velocities can vary as a result of differences in the diameters of the muscle fibers. Consequently, the potentials generated by different muscle fibers of the same MU have temporal dispersion. Electrodes inserted inside the territory of a discharging MU detect these potentials within its uptake area [12].

Single muscle fiber generated waveforms are called muscle fiber potentials (MFPs) and the summating potentials generated by different muscle fibers of the same motor unit are called motor unit potentials (MUPs). To maintain or increase the exerted force produced by a muscle, MUs are repeatedly activated and thus generate trains of MUPs which are called motor unit potential trains (MUPTs). MU firing frequency is defined as the number of recurring MUPs along a train per second. Note that the greater the number of MUs activated and their discharge frequency, the greater the generated force will be. Finally, an EMG signal is defined as the summation of the MUPTs detected during a contraction [7].

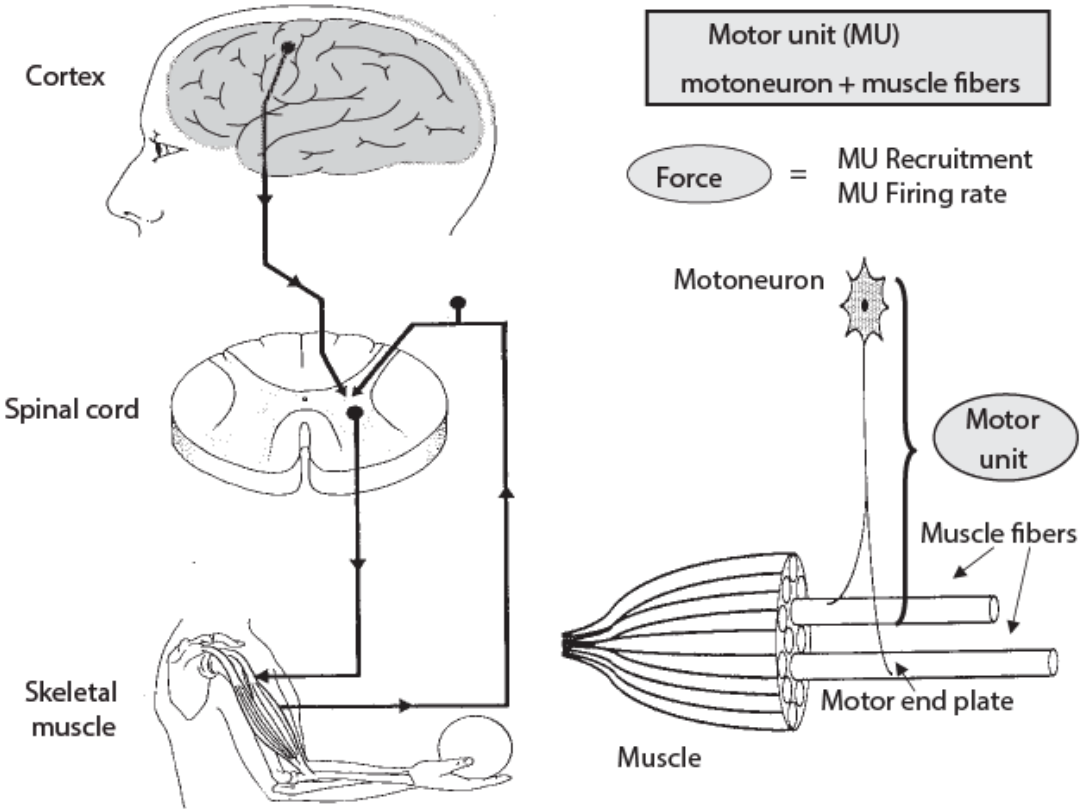


Figure 2.1: A simple representation of the motor unit and motor control process [2].

EMG signals can be detected using two possible types of electrodes: needle and surface

electrodes. For clinical applications needle electrodes are utilized to record EMG signals from the muscles despite their invasive nature and the discomfort they may cause patients. A needle electrode is inserted directly into the muscle and is well suited for inferring motor unit internal structure and size properties. As an example, fibrillation potentials, which are potentials generated by single denervated muscle fibers and an important sign of disease can only be detected using needle electrodes. On the other hand, signals recorded using surface electrodes are better for analyzing the temporal pattern of activity, and fatigue of a muscle as a whole or of muscle groups. They are mostly used in sport, rehabilitation and occupational medicine where assessments have to be repeated frequently [2].

Neuromuscular disorders are a heterogeneous group of diseases impairing muscle function. Two main categories of neuromuscular disorders are myopathic and neurogenic where the former is caused by the atrophy or death of muscle fibers and the latter is primarily caused by damage to or death of motor neurons. The primary symptom in myopathies is muscle weakness which is the result of dysfunctional muscle fibers. Other symptoms can be muscle cramps, stiffness, and spasm [8].

In neuropathies, loss of motor neurons is an early sign, where fibers associated with degenerating motor neurons, lose their neuronal connection and become denervated. Hence, surviving motor neurons in the immediate vicinity of the orphaned/denervated muscle fibers grow new axonal sprouts and re-innervate the denervated muscle fibers. Neuropathies have different symptoms. Some patients may have temporary numbness, tingling, and pricking sensations, sensitivity to touch, or muscle weakness. Others may experience more severe symptoms, such as burning pain (especially at night), muscle wasting, paralysis, or organ or gland dysfunction [13].

2.1.2 MUPT Extraction Using EMG Decomposition

EMG signals are defined as the linear summation of the MUPTs generated by the MUs active in an examined muscle. Individual MUPTs can be extracted and analyzed from EMG signals to assist in the diagnosis of neuromuscular disorders. The main purpose of

EMG signal decomposition is to extract significant, constituent MUPTs from the composite EMG signal. To this end, the decomposition algorithm should have the ability to identify MUPs of the MUs significantly contributing to the composite EMG signal and to correctly associate each detected MUP with the MU that generated it. Fig. 2.2 shows EMG signal decomposition process conceptually and represents the relationship between a decomposed EMG signal and the activity of individual MUs. EMG signal decomposition involves two or three main steps which are explained as follows [14], [15]:

1. EMG Signal Segmentation: The first step of EMG signal decomposition is signal segmentation. The main purpose of this step is to segment the EMG signal into sections which contain significant MUPs. To this end, some detection thresholds are defined according to some statistic calculated using the composite EMG signal. When the signal characteristics generate a statistic value above the threshold value, a fixed or variable length section(s) are selected which are assumed to contain significant MUP contributions. Detection thresholds are usually based on signal amplitude, slope, or both [16].

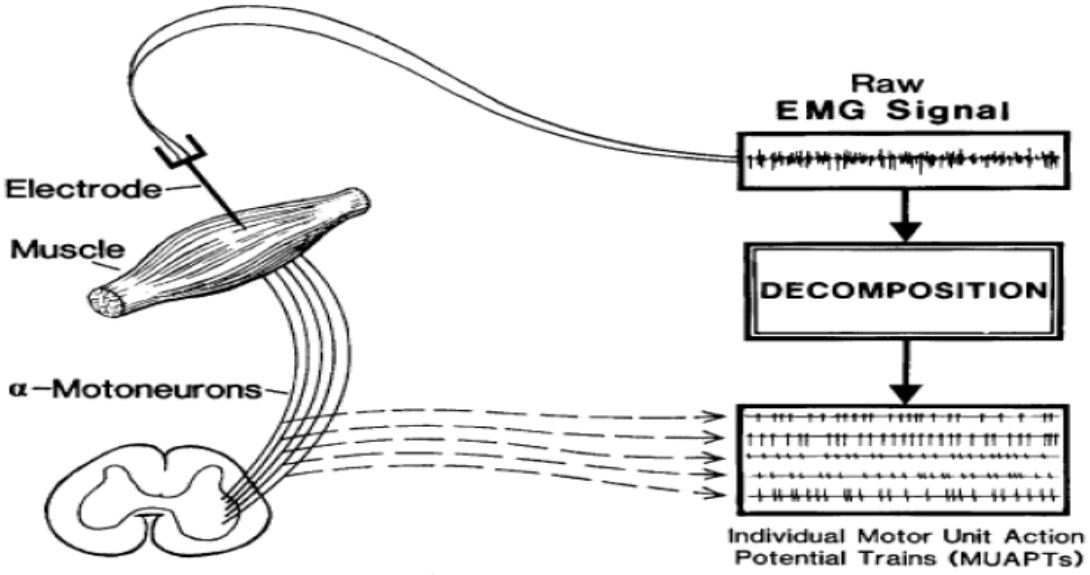


Figure 2.2: MUPT can be extracted using EMG decomposition technique. Here, five motor units were detected and their MUPTs were extracted using an EMG decomposition program [3].

2. MUPs Clustering: After the MUPs are extracted from the composite EMG signal, the correct number of significant MUPTs should be estimated. In addition, the MUPT template shape of each contributing MU should be determined. These two tasks can be performed by clustering the MUPs contributed to the whole EMG signal. MUPs clustering partitions MUPs into a number of clusters. MUPs in the same cluster should be more similar to the other members of its cluster than it is to the MUPs of any other cluster. Euclidean distance, a dissimilarity measure, along with different clustering techniques such as hierarchical, partitioning and density-based clustering are often employed for this purpose.

Each of the obtained clusters represents a MUPT and a MUPT template is defined as the mean of the characteristics of all the MUPs belonging to the same cluster. Some of the existing decomposition algorithms are designed only based on clustering. In some others, clustering is performed only on the MUPs contributed to the initial t second of the signal and then a classifier is trained by the information obtained in the clustering phase. In these cases, the objective of clustering is to provide the necessary initial information required for classification such as the number of classes or MUPTs, a prototype for each MUPT, and the motor unit firing pattern statistics for each MUPT [17], [18].

3. MUP Classification: Some decomposition algorithms are designed only based on clustering [19], [20], [21]; however, in most algorithms, the clustering algorithm is followed by MUP classification [22], [23], [24], [25] and the classification procedure is repeated across several number of iterations. These iterations are terminated when the extracted MUPTs are stable or some termination criteria are satisfied. As such, first, the MUP template and MU firing pattern statistics of each MUPT are estimated based on the results of the clustering stage. Next, MUP shape similarity assignment thresholds for each MUPT are calculated. Finally, MUPs are classified to the extracted MUPTs with regards to their shape and MU firing pattern statistics.

In some algorithms, MUP template and MU firing pattern statistics of each MUPT are updated after classification. In addition, similar MUPTs are merged if they represent

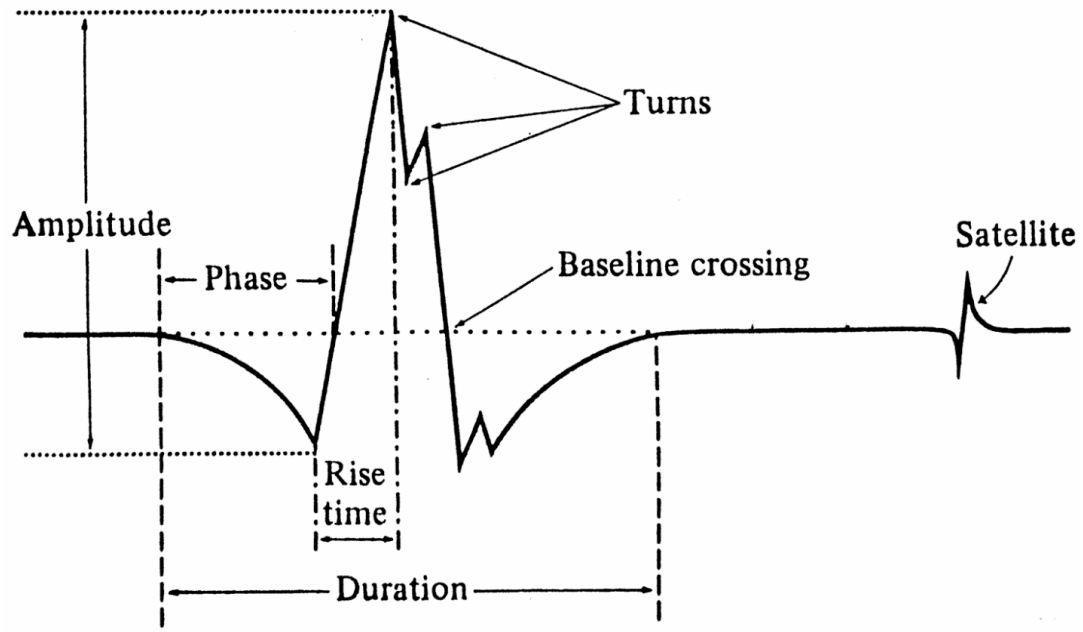


Figure 2.3: Morphological parameters of a motor unit potential [4].

the activity of the same MU, and then a new iteration for MUP classification is repeated based on the updated MUP template and MU firing pattern statistics. Once MUPTs are extracted from EMG signals using a decomposition algorithm, various features (see Fig. 2.3) can also be obtained by applying quantitative techniques. These features can help to determine whether the muscle is affected by a disease process and, if so, whether it is mild or sever.

2.2 Related Work

2.2.1 Critical Review of the Existing Electrophysiological Muscle Classification Techniques

Electrodiagnosis of most neuromuscular disorders is based on the classification of involved muscles. Table 2.1 presents details of some representative quantitative EMC systems. Quantitative EMC techniques can be categorized as either muscle or MU level methods based on the space or level in which the discriminative information used lies. Muscle-level methods treat each muscle as a whole entity and train a classifier directly on discriminative muscle level information extracted from acquired composite EMG signals [26] whereas, MU-level methods classify muscles based on the aggregation of MU-level characterization scores of a set of sampled MUs [2].

Muscle-level methods, sometimes called "interference pattern analysis", may not provide sufficient sensitivity for clinical application. This limitation is based on superpositions of MUPTs which makes detection of marginal levels of disease involvement difficult. Hence, small numbers of abnormal MUPTs may be obscured in composite EMG signals generated by a majority of normal MUPTs. As a result, for clinical application, MU-level methods are usually preferred.

MU-level muscle classifications are usually performed as a two-step procedure starting with the characterization of a set of extracted MUPs followed by the aggregation of the MUP characterizations. MUP characterization, in turn, is performed based on two different approaches. In the first approach, a MUP is classified by determining if it was produced by a normal or diseased MU. For this approach, an expert physician labels individual MUPs in the training data and such labeling may be prone to errors and may not be feasible for datasets comprised of a large number of extracted MUPs [19].

There are several reported supervised learning methods for determining MUP characterization labels. Artificial neural networks (ANNs) have been used extensively for MUP

Table 2.1:
Summary of the Representative Electrophysiological Muscle Classification Systems.

ID	Classifier	Year	Muscle	#of Muscles	#of MUPs	MUP Labelling	Discriminative Information Level	Accuracy(%)
1	ANNs	1995 [2]	Biceps Brachii	44	880	Physician	MU-Level	80-90
2	ANNs	1996 [27]	Right Biceps	50	-	-	Muscle-Level	60-80
3	ANNs	1998 [28]	Biceps Brachii	40	800	Physician	MU-Level	79.6
4	ANNs	2007 [29]	Biceps Brachii	62	365	Physician	MU-Level	89
5	SVMs	2005 [30]	Biceps-Hypothenar Group	59	-	-	Muscle-Level	92.3
6	SVMs	2010 [31]	Biceps Brachii	27	-	-	Muscle-Level	70.4
7	SVMs	2012 [32]	Biceps Brachii	27	-	-	Muscle-Level	97.67
8	SVMs	2013 [33]	Biceps Brachii	27	-	-	Muscle-level	96.75
9	Fuzzy	2012 [26]	Biceps Brachii	27	-	-	Muscle-Level	93.3
10	DTs	2012 [34]	Biceps Brachii	27	-	-	Muscle-Level	96.05
11	RFs	2015 [35]	Biceps Brachii-Medial Vastus	25	-	-	Muscle-Level	96.67
12	KNN	2014 [36]	Biceps Brachii-Medial Vastus	25	-	Muscle-Label	MU-Level	98.8
13	GMMs	2014 [37]	Tibialis Anterior-Fiurst Dorsal Interosseous-Deltoid- Vastus Medialis	342	5764	Probabilistic Labels	MU-Level	88.17

classification due to their simplicity and ability to model complex non-linear systems. These studies have found that ANNs have a high tendency towards overfitting and do not generalize well. In addition, ANNs can be categorized as black box classifiers and do not provide neccessary transparency due to the large number of transformations applied to the input feature vectors [2], [20], [27], [28].

Linear discriminant analysis (LDA) has been used by several researches for MUP classification. LDA works best in cases having features with continuous quantities, and has been found to be useful for MUP characterization. However, there are serious limitations with LDA. LDA does not work well if the dataset is not balanced such that the number of objects in various classes is highly different. In addition, LDA is not applicable for separation of non-linear problems [32], [38].

Fuzzy logic techniques have been used for MUP classification with the purpose of improving the transparency of MUP labelling results [26], [39]; however accuracy may decrease. In addition, constructing rules from domain knowledge is a tough task for human experts especially in cases having feature spaces of more than three dimensions. Decision trees (DTs) have also been used with the purpose of increasing transparency while maintaining high accuracy [20], [34] however the greedy nature of DTs can lead to a high susceptibility to outliers and a tendency towards overfitting.

Support vector machines (SVMs) have been used in a variety of studies [30], [40], [41], [42], [43] for MUP classification due to their high generalization capability. In addition, an SVM can discriminate classes that have nonlinear complex decision boundaries. In most of these works, first, an SVM classifier was used to discriminate the healthy subjects from the diseased ones and then another SVM classifier was used to classify the diseased subjects into myopathic and neurogenic classes.

Ensemble methods have been used for MUP classification [44], [45]. Ensemble models build a strong learner from a group of weak learners with the purpose of increasing accuracy while preventing overfitting. To this end, opinions of multiple learners are combined by following several aggregation schemes such as majority voting and weighted voting.

In the second approach to MUP characterization, MUPs in the training data are initially labeled based on the clinical classification of their corresponding muscle (i.e., MUPs belonging to a myopathic muscle are all labeled as myopathic). A learning algorithm is then applied to estimate a set of likelihood scores, one each for the MUP being detected in a normal, myopathic and neurogenic muscle, respectively [29]. For this approach, the MUP characterization scores are not based on the likelihood of the MU that generated the MUP being normal, myopathic or neurogenic, but rather the likelihood of it being detected in a muscle of a specific class. In addition, normal and diseased muscles have normal MUs which in turn cause MUP characterization score distributions to be highly overlapped. Therefore, because this method of MUP characterization does not directly reflect the presence of disease, the accuracy with which the obtained MUP characterization scores can be used to classify muscles suffers.

There are several reported supervised learning methods for determining MUP characterization scores. Pattern discovery (PD) which is based on quantization of all continuous feature values into discrete events was used for MUP characterization. The main limitation with PD, compared to other characterization methods, is a possible lack of robustness due to the need to discretize continuous feature values [46]. Gaussian mixture models (GMMs) have also been used [37], but are susceptible to outliers because of their reliance on maximization of likelihood functions assumed to have Gaussian distributions.

For muscle classification, obtained MUP characterization labels or scores need to be aggregated. In previous works, the aggregation of MUP characterization scores was accomplished using fixed combination rules (ex. weighted sum and maximum) or a Bayesian aggregation rule [46], [37]. Fixed combination rules are sensitive to extreme values and cannot provide accurate measures in cases with extreme values or dispersed MUP characterization scores. In addition, muscle classification scores generated using Bayes rule tend to saturate to 0 or 1 as more evidence (higher number of diseased MUPs) is presented, which is not desirable when the classifier is uncertain of the outcome. In this case, the classification score provided should reflect the level of uncertainty, rather than saturate to a value supporting an incorrect category [46].

2.2.2 Existing Multiple Instance Learning Techniques

A classifier in the standard supervised learning paradigm is designed according to a training set consisting of instances with associated class labels. In contrast, a classifier in the multiple instance learning (MIL) paradigm is designed according to a training set consisting of bags of instances, where each bag has an associated label, but the individual instances usually do not have any label. In addition, all the instances in the bag are not necessarily relevant and there might be some instances inside one bag that do not convey any information about the bag class label, or that are more relevant to other classes of bags, providing confusing information [9].

A bag is a set, and the elements of the set are instance feature vectors, and the number of instances inside bags may be different across the training set. All the instances are represented in a d-dimensional feature space, called the instance space. The main goal of the MIL classification problem is to learn a model which can be used to predict the class label of unseen bags. The usual framework in the MIL literature is binary classification (positive vs. negative instances), hence for multi-class problems, a one-versus-all strategy is usually employed [47]. The MIL classification methods are broadly categorized into two groups based on how the multiple instance information is exploited. These categories are usually called instance-level and bag-level MIL paradigms [48], [49].

In the instance-level paradigm, the discriminative information is exploited at the instance level. Hence, an instance-level classifier is trained to classify the instances. The bag level classifier is obtained by aggregating instance level scores generated by the instance-level classifier. The key challenge in instance-level methods is how to infer an instance level classifier without having a training set of labelled instances [50], [51]. To this end, some assumptions must be made about the relationship between the labels of the bags in the training set and the labels of the instances contained in these bags [9]. In this sense, two sub-categories of instance-level methods emerge in the literature: the ones following the classical multiple instance assumption and the ones following the collective assumption. The classical multiple instance assumption states that only a small fraction of the instances

inside the bag provide information about the class label of the bag. A well-known example of the classical multiple instance assumption is called the standard assumption which assumes every positive bag contains at least one positive instance, while all the instances in every negative bag are negative. In contrast, the collective assumption states that all instances in a bag contribute equally to the bags label [50], [52].

In the bag-level paradigm, the discriminative information is exploited at the bag-level. Hence, each bag is treated as a whole entity and a bag-level classifier is trained to classify the whole bag. Some bag-level methods assume that bags from the same class are similar. In most implementations, all the instances of the bags are involved in calculating bag similarity. Based on the obtained similarities, a bag-level classifier is trained to predict the label of unseen bags. These methods usually do not assign labels to the instances and the classifiers are trained to classify the whole bags [53], [54].

Chapter 3

Problem Formulation

In this chapter, the problem of electrophysiological muscle classification is mathematically formulated and an overview of the proposed EMC system is presented.

3.1 Overview of the Electrophysiological Muscle Classification Problem

Physicians diagnose neuromuscular disorders by considering characteristics of a set of examined muscles. An individual muscle is electrophysiologically characterized as normal, myopathic or neurogenic, by first qualitatively analyzing sets of MUPs representing MUs sampled in the examined muscle and then aggregating their MUP characterizations. Quantitative EMC systems, like physicians, consider sets of MUPs representing sampled MUs, and quantitatively characterize individual MUPs and then aggregate these characterizations to obtain a muscle classification. Fig. 3.1 shows a schematic representation of a quantitative EMC system.

In electrophysiological muscle data, training examples are not singletons. Instead, they are presented in the form of bags of instances where each bag represents a muscle and

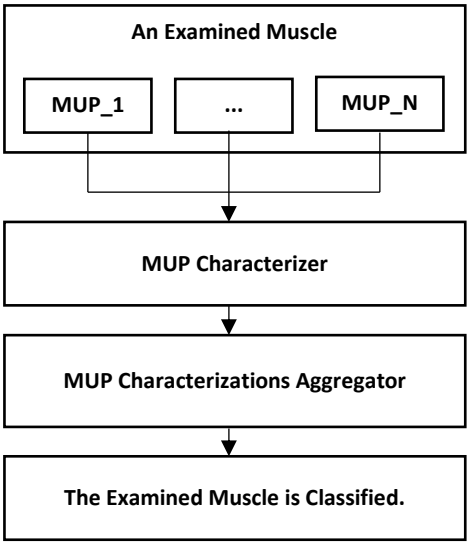


Figure 3.1: A simple schematic representation of a typical quantitative EMC system

has an associated label, and the instances inside a bag are MUPs extracted from needle-detected EMG signals acquired from that muscle. The individual instances do not have any label. All the instances inside the bag are not relevant to the class label of their muscle and they may provide confusing information. Furthermore, the instances inside one bag may even be more related to a class other than the class of their muscle. The number of instances inside each bag is not the same, however they all lie in a common d-dimensional feature space.

Considering the above muscle data characteristics, it is difficult to formulate the muscle classification problem in a standard supervised learning setting in which both training and test data are represented as individual feature vectors. Instead, the training and test data can be represented by sets or bags of feature vectors or instances. In other words, the electrophysiological muscle data consists of muscles (bags) with associated labels and each electrophysiological muscle data set consists of several MUPs (instances) that do not have any labels. Furthermore, the muscle-level classifier should by design be induced by a MUP-level classifier. Consequently, the problem of electrophysiological muscle classification can

be naturally cast in an MIL setting, which unlike supervised learning, does not require label information for each training instance, but rather for collections of instances or bags.

Mathematically, the MIL-based electrophysiological muscle classification (MIL-based EMC) problem can be presented in the following form. The i $^{\text{th}}$ muscle has a label $Y_i \in \Upsilon$, where Υ denotes the set of possible class labels for muscles. A muscle (bag) is represented by the set of MUPs (instances) extracted from the clinically-detected EMG signals acquired from the muscle. Let $X_i = \{\overrightarrow{x_{i1}}, \ldots, \overrightarrow{x_{in_i}}\}$ denote the set of n_i MUPs detected in the i $^{\text{th}}$ muscle. Each MUP is represented by an M dimensional feature vector $\overrightarrow{x_{ij}} \in R^M$. The muscle classification training set is then represented as $D = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ where N is the number of muscles used to create the dataset. Given this, the final objective is to learn an MIL-based electrophysiological muscle classifier, which classifies a muscle as being healthy or affected by a disease process. If the muscle is diseased, the MIL-based electrophysiological muscle classifier should also determine the type of disease.

3.2 Overview of the Proposed Method

For the purpose of diagnosing a neuromuscular disease process and determining its level of involvement, MUPs extracted from the EMG signals detected from the examined muscle should be characterized. In other words, several classes of MUPs presented in the muscle data need to be identified. These classes are discovered in an unsupervised way and can be obtained by running a clustering algorithm on the MUP feature space. Based on the obtained classes, the MUPs of the examined muscle are characterized. The examined muscle, then, can be classified with regards to the aggregation of its MUP characterization classes.

Current quantitative EMC systems are not sufficiently accurate and robust to be reliably used for clinical needle electrodiagnostic examination. They consider only three possible classes for both MUP characterization and electrophysiological muscle classification. Because disease causes continuous, rather than discrete, modifications to both MU

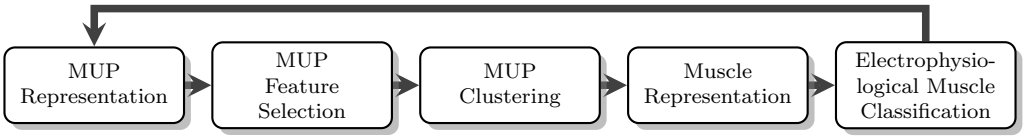


Figure 3.2: Steps of the proposed MIL-based electrophysiological muscle classification system

fiber spatial and diameter distributions, disease induced changes in MUP characteristics will also be continuous in nature. This means that, except for extreme cases, it is not possible to find distinct boundaries between MUPs generated by normal, myopathic or neurogenic MUs. As such, a MUP characterized based on the type and severity of the effects of a disease process on its generating MU and not just three conventional groups, is preferred.

Discovering a dynamic number of distinct groups in a set of MUPs represented in a MUP feature space requires a clustering algorithm capable of finding natural clusters. A natural cluster has arbitrary shape and density, a fact not considered in several groups of clustering algorithms as they assume clusters to have globular shape or a specific user defined density. However, disease-induced MUP changes are continuous and possibly nonspecific in nature, which makes it difficult to find distinct normal versus myopathic versus neurogenic class boundaries. Hence, MUP characterization based on more than three classes may better represent the various effects of disease.

There are three main groups of clustering algorithms which can be categorized based on their ability for finding natural clusters. The first group includes those algorithms that can find clusters with specific shape and specific density (SSSD). For instance, center-based clustering algorithms like k-means [55] assume clusters to have known geometrical shapes like a sphere or an ellipse. A second group of clustering algorithms includes those that can find clusters with arbitrary shape and specific density (ASAD). As an example, some density-based clustering algorithms, including DBSCAN [56], use a static model to characterize the density of the data. A third group of clustering algorithms includes those that

are able to find clusters with arbitrary shape and density(ASAD). These algorithms usually employ local and global similarity information to organize data into sensible groupings with dynamic properties such as density and shape. As an example, Chameleon [57], a hierarchical clustering algorithm, uses a dynamic global model to find clusters with arbitrary shape and density. With Chameleon, the creation of several initial clusters and the overhead of using graph partitioning leads to increased time complexity.

In this thesis, a novel EMC system is proposed that focuses on using a dynamic number of classes for characterizing MUPs. To this end, a novel density-based clustering algorithm called Neighborhood Distance Entropy Consistency (NDEC) is proposed to find representations of several concepts of normality and abnormality in the MUP feature space to be used for MUP characterization. The MUPs sampled from an examined muscle are then characterized, using the increased number of MUP classes, and the MUP characterizations are embedded into a feature vector to represent the examined muscle. The embedded feature vector is then fed to an appropriate supervised classifier to obtain the electrophysiological muscle classification. Fig. 3.2 shows the main steps of the proposed MIL-based EMC system.

Chapter 4

NDEC: A Density-Based Clustering Algorithm using Neighbourhood Distance Entropy Consistency

The work described in this chapter previously appeared in:

T. Kamali and D. W. Stashuk, NDEC: A Density-Based Clustering Algorithm using Neighbourhood Distance Entropy Consistency, IEEE Transactions on Knowledge and Data Engineering, 2018 (Under Review).

T. Kamali and D. W. Stashuk, Automated Segmentation of White Matter Fiber Bundles using Diffusion Tensor Imaging Data and a New Density Based Clustering Algorithm, Artificial Intelligence in Medicine 73 (2016): 14-22 [58].

4.1 Introduction

Electrophysiological muscle classification training data is an example of a multiple instance learning paradigm and consists of data arranged in collections called bags. In this paradigm, a label is associated with each bag, however the individual instances inside the bag do not have any label. Furthermore, all the instances in the bag are not necessarily relevant and there might be some instances inside one bag that do not convey any information about the bag class label, or that are more relevant to other classes of bags, providing confusing information. In the framework proposed in Chapter 3 for electrophysiological muscle classification, we are interested in clustering the instances of all bags (muscles). This task is totally unsupervised and aims to identify classes of instances that are present in muscle training datasets. The obtained clusters will next be used as instance (MUP) characterization classes and will be utilized to construct a more discriminative feature set for a subsequent muscle classification procedure. These clusters will have an associated semantic label based on expert domain knowledge. In this chapter, a new density-based clustering algorithm is proposed which is used to find a dynamic number of MUP characterization classes.

One of the most fundamental tasks in data mining is clustering. Cluster analysis refers to the unsupervised classification of patterns into groups or clusters based on their similarity [59]. This has applications in a wide variety of fields ranging from geographic information systems, to bioinformatics, image segmentation, and pattern recognition [60], [61], [62], [63]. Clustering can be used as an efficient tool to observe hidden patterns and meaningful concepts in an analyzed dataset [64]. A literature survey shows that no established definition of a cluster exists and the definition of a cluster is usually dependent on the nature of the dataset under study. Therefore, there exists a wide range of clustering algorithms with diverse assumptions about the structure of a given data set. With respect to generalization, a clustering algorithm which uses few assumptions about the density, shape and structure of the obtained cluster is highly preferred.

Finding natural clusters with arbitrary shape, size, and density is challenging due to

the complex nature of the problem. The main challenges of this task are briefly explained below:

1. Little/No Prior Information: Most clustering algorithms find clusters based on prior assumptions. For instance, some algorithms assume that the number of clusters is known in advance and that the shape of all the clusters are convex (i.e. globular or spherical). Other algorithms assume that the density of each cluster is similar. These assumptions restrict the ability of the clustering algorithm to find natural clusters and result in poor performance when the clustering assumptions are violated [65].

2. Expensive Similarity Measures: Most clustering algorithms are designed based on capturing similarity between data points. Many similarity measures can be computed between data points. Most of these measures have high computational complexity. Hence, calculating all pair-wise similarity measures is not computationally efficient and/or feasible in a variety of applications [66].

3. Clustering Result Validation: The process of estimating how well a partition fits the structure underlying the data is known as cluster validation. Usually there is no ground truth knowledge available about the data structure, hence, determining the utility of clustering results is challenging [67].

4. Identify Noise: An important aspect of clustering is how to identify noise objects. Noise is defined as those objects that do not belong to any cluster. Given this definition, an ideal clustering algorithm is the one which can correctly assign samples to their corresponding clusters and delineate samples that do not belong to any clusters.

4.2 Related Work

A good cluster analysis should capture the intrinsic structure of the data and be able to identify clusters that have arbitrary shape and density. This principle is not considered by several groups of clustering algorithms as they assume clusters to have globular shape

or a specific user defined density [68]. Clustering algorithms can be divided into three main groups with respect to their abilities for finding natural clusters. These groups are explained briefly below.

1. Specific Shape and Specific Density (SSSD): Clustering algorithms in this group can find clusters with specific shape and specific density. As an example, center-based clustering algorithms like k-means [55], PAM [69], CLARANS [70] and BIRCH [71] assume clusters to have known geometrical shapes like a sphere or an ellipse. In addition, some hierarchical clustering algorithms including Single, Average and Complete linkage [72], and CURE [73] fail to find clusters with arbitrary shapes.

2. Arbitrary Shape and Specific Density (ASSD): Clustering algorithms in this group can find clusters with arbitrary shape but specific density. For instance, grid-based models like Wave Cluster [74] and some density-based clustering algorithms including DBSCAN [56], DENCLUE [75], SSN [76] and GDD [77] use a uniform model to characterize the density of the data. Some hierarchical clustering algorithms, such as OPTICS [78], propose methods to discover a flat partition by utilizing a global density threshold. OPTICS may not be able to find the most significant clusters if these clusters have different density [79]. The clustering algorithms in the ASSD group rely on user defined parameters to determine constant density thresholds and do not consider the actual density distributions of the underlying data. As a result, their abilities are limited to finding clusters with arbitrary shapes but constant densities.

3. Arbitrary Shape and Arbitrary Density (ASAD): Clustering algorithms in this group are able to find clusters with arbitrary shape and density. In ASAD group, usually local and global similarity information are utilized to organize data into meaningful/useful groupings with dynamic shape and density properties [80].

In early work, Zahn [81], represented each cluster using a minimum spanning tree (MST) and removed inconsistent edges from the MST to obtain a clustering solution. According to Zahn's definition, an edge in a MST is inconsistent if its weight is larger than its neighbouring edge weights mean and standard deviation considering a specified

factor. This approach only works for simple data sets since it only measures local density consistency.

In contrast, a hierarchical agglomerative clustering algorithm called Chameleon [57] finds clusters with arbitrary densities and shapes based on a dynamic global model. This model utilizes the k-nearest neighbours (kNN) graph, relative inter-connectivity and closeness indices weighted by user defined parameters to discover the clusters. Chameleon is robust to noise and outliers, however, it has problems when the partitioning process does not generate sub-clusters which happens often when clustering high-dimensional data. In addition, for testing each individual parameter, a complete hierarchy needs to be created which makes parameter selection difficult and time consuming. Furthermore, Chameleon only considers the consistency of edges in a global sense [82].

Another example of a dynamic density based clustering algorithm is Mitosis [83], which attempts to find clusters with arbitrary shape and density based on distance-relatedness concepts. Initially, Mitosis creates associations for each individual pattern in the data based on local distance similarity and then puts associated patterns into one cluster if they satisfy distance consistency criterion. For singleton patterns in associations, local distance relatedness information is considered, whereas for non-singletons global distance relatedness information is evaluated. Mitosis tries to overcome some of the weaknesses of Chameleon by employing either local or global distance information in each merge step. However, as cluster sizes increase, given that some clusters have non-homogeneous density, the algorithm fails to find appropriate clusters.

In density peaks clustering (DPC)[84], the center of a cluster is identified as the sample which has the highest local density among its neighbor points. In addition, the center of a cluster has a relatively large distance from other points with higher densities. These two criteria are utilized to determine the cluster centres. The clusters are then constructed by assigning points to the same cluster to which its nearest neighbor of higher density belongs. DPC provides an effective and simple way to discover clusters with arbitrary shape and density in most cases. However, when a cluster has more than one center, DPC may fail

to find the most significant clusters.

4.3 The NDEC Clustering Algorithm

Algorithm 1 shows the pseudocode of the NDEC clustering algorithm. NDEC attempts to find natural clusters with arbitrary shape and density without utilizing any prior knowledge or assumptions about the number, shape and density of the clusters. To this end, the following steps are performed. 1) Local density estimation using the k-nearest neighbours; 2) Generation of sub-clusters based on local and global density consistency; 3) Generation of final clusters based on neighbourhood distance entropy consistency; 4) Outlier identification and handling. Details of each step are explained below.

4.3.1 Preprocessing

In this step, local density is estimated using the k-nearest neighbors and an abstraction for the dataset under study is created. According to Hartigan [55], clusters are identified as regions with high density isolated from other such regions by regions of low density. Density can be estimated using parametric or non-parametric methods. Parametric methods simply estimate parameter values of an assumed distribution shape which can restrict the adaptivity of the density estimates to intrinsic data characteristic [85], which in turn restricts the discovery of clusters with arbitrary properties. Non-parametric methods do not suffer from these restrictions.

Clusters with arbitrary shape and density have varying local densities. This issue can be resolved by determining the density of a region based on the behaviour of the neighbourhood distances. In this scenario, smaller distances relate to higher densities and larger distances relate to lower densities. Consequently, in NDEC local density is estimated based on local averaging of nearest neighbour distances which are determined using a binary metric tree similar to the one proposed in [86] to improve the efficiency of searching through

high dimensional data.

Local Density Information (LDI): Let p be an arbitrary sample in the dataset D, and $\mathrm{NN_k(p)}$ be a set that consists of the k-nearest neighbour distances of p where k is a user defined parameter. The LDI for p is then defined as the average of the distances ($\mathrm{d_i}$) in $\mathrm{NN_k(p)}$.

$$LDI(p) = \frac{1}{k}\sum_{i=1}^{k}\{d_{\mathrm{i}} \in NN_{\mathrm{k}}(p)\} \tag{4.1}$$

The information about the nearest neighbours of individual samples is saved in the $list_{NN}$ (see Eq. 4.2), where p and q are two arbitrary samples in D, and $d\ (p,\ q)$ is the symmetric Euclidean distance between them. After $list_{NN}$ is created, it is sorted ascendingly based on $d\ (p,\ q)$ and duplicate associations are removed from it. From this point forward, $\mathrm{list_{NN}}$ is used as an abstraction for dataset D.

$$list_{\mathrm{NN}} = \{(p, q, d\,(p, q)\ )|\ p, q\ \in D\ \ \wedge q\ \in NN_k\,(p)\}_{\,\forall p \in D} \tag{4.2}$$

4.3.2 Generation of Sub-Clusters based on Local and Global Density Consistency

NDEC creates clusters following an agglomerative hierarchical approach. In this way, $list_{NN}$ is traversed and eligible members of this list are put into clusters. This eligibility is initially specified by considering the consistency of local density information and then after creating sub-clusters, the consistency of global density information is considered. The NDEC clustering algorithm starts merging singletons and creating clusters by traversing through $list_{NN}$. Each time a tuple $(p,q,d(p,q))$ is analyzed, if $d(p,q)$ is consistent with both $LDI(p)$ and $LDI(q)$ and simultaneously $LDI(p)$ and $LDI(q)$ are consistent (see Algorithm 1), then p and q are put into one cluster and their associated tuple is removed from $list_{NN}$.

While traversing through $list_{NN}$ and analyzing an association *(p,q,d(p,q))*, if either of p or q belongs to a cluster, global density information (GDI) is calculated and used to define the distance consistency criteria (see Algorithm 1) and if the distance consistency criteria are met for a singleton sample, it is assigned to the cluster. The GDI of the cluster C_p is defined based on the following definition.

Global Density Information (GDI): Assume C_p represents the cluster to which sample p belongs, $list_{NN}(C_p)$ is a list of all associations that belong to C_p, and N_p is the length of $list_{NN}(C_p)$, the $GDI(C_p)$ is calculated as the average of the distances (d_j) in $list_{NN}(C_p)$:

$$GDI\left(C_p\right) = \frac{\sum\limits_{\forall d_j \in list_{NN}(C_p)} d_j}{N_p} \tag{4.3}$$

4.3.3 Generation of Final Clusters based on Entropy Consistency

Entropy is an index that measures the amount of irregularity within a set of data. The consistency of the neighbourhood distance entropy of two sub-clusters implies density homogeneity within the potentially merged cluster. Consequently, while traversing through $list_{NN}$ and analyzing an association *(p,q,d(p,q))*, if both p and q belong to cluster(s), a new measure, called "Global Entropy Information (GEI)" is calculated and if the GEIs are consistent the candidate sub-clusters are merged.

The entropy *H(f)* of a continuous probability density *f(x)* is calculated by the following equation.

$$H\left(f\right) = -\int_{-\infty}^{\infty} f\left(x\right) log f\left(x\right) dx \tag{4.4}$$

This method of calculating entropy can be used when *f(x)* is known. However, in some cases the density *f(x)* is unknown, but a sample of size N from this density is available.

In this case, two approaches can be followed to calculate entropy. The first approach is called a plug-in estimate in which any standard density estimation technique is first used to estimate the unknown density *f(x)* from a sample and then, the entropy of the density estimate $\hat{f}(x)$ is calculated as an estimate of the true entropy of f [87]. Plug-in estimates is well suited for densities with known parametric form.

There is another method which is used for estimating one-dimensional entropies. This method is based on the order statistics of a sample and provides a consistent and rapidly converging entropy estimator which can be used when the density *f(x)* in unknown. As a result, in this work, entropy estimation based on order statistics is used [88].

Global Entropy Information (GEI): Assume C_p represents the cluster to which sample p belongs, $list_{NN}(C_p)$ is the list of all associations that belong to C_p and N_p is the length of $list_{NN}(C_p)$, then the $GEI\ (C_p)$ is computed using the following steps:

1. Calculate the order statistics of the distances in $list_{NN}(C_p)$, which is simply the distances of $\text{list}_{NN}(C_p)$ rearranged in an increasing order $\left(\left\{d^{(1)},\ d^{(2)},\ \ldots,\ d^{(N_p)}\right\}\right)$.

2. Estimate the entropy of the distances in $list_{NN}(C_p)$ using Eq. 4.5 based on calculating the m-spacings of the order statistics of the distances in $list_{NN}(C_p)$, where $m = \sqrt{N_p}$. This entropy estimator is asymptotically efficient and was proposed by Miller [89].

$$GEI\,(C_p) = \frac{1}{N_p - m} \sum_{n=1}^{N_p - m} Log(\frac{N_p + 1}{m}\ (d^{(n+m)} - d^{(n)})) \tag{4.5}$$

4.3.4 Outlier Identification and Handling

There is no standard procedure in the literature explaining how to handle noise objects. Usually based on the problem domain, one of the following alternatives is employed: 1) assign each noise sample to a singleton cluster, 2) assign all noise samples to a single cluster, 3) discard all noise samples, and 4) assign each noise sample to its closest cluster

Algorithm 1: NDEC

Input : Dataset D with n samples, k (number of nearest neighbors), l (distance consistency) and h (entropy consistency) user-defined parameters

Output: A set of discovered clusters and outliers.

NN(i): A list of k nearest neighbors of sample i
$List_{\mathrm{NN}} \leftarrow \emptyset$

for *i=1:n* **do**
    $\mathrm{List_{NN}}$.append(i, NN(i), dist(i, NN(i));
    $\mathrm{List_{NN}}$.RemoveDuplicates;
    $\mathrm{List_{NN}}$.Sort('ascend');
end

while *Change ($List_{NN}$.Length)* **do**
    for *i=1:$List_{NN}$.Length* **do**
        (p,q,d(p,q))=$\mathrm{List_{NN}}$(i);
        if $Singleton(p) == True \wedge Singleton(q) == True$ **then**
            if $d(p,q) < l \times Min(LDI(p), LDI(q)) \wedge$
            $Max(LDI(p),LDI(q)) < l \times Min(LDI(p), LDI(q))$ **then**
                Put p and q into one cluster
                $\mathrm{List_{NN}}$(i).Delete;
            end
        else if $Singleton(p) == True \wedge Singleton(q) == False$ **then**
            if $d(p,q) < l \times Min(LDI(p), LDI(q)) \wedge$
            $Max(LDI(p),GDI(q)) < l \times Min(LDI(p), GDI(q))$ **then**
                Put pattern p into cluster associated with q
                $\mathrm{List_{NN}}$(i).Delete;
            end
        else if $Singleton(p)==False \wedge Singleton(q)==True$ **then**
            if $d(p,q) < l \times Min(LDI(p), LDI(q)) \wedge$
            $Max(GDI(p),LDI(q)) < l \times Min(GDI(p), LDI(q))$ **then**
                Put pattern q into cluster associated with p
                $\mathrm{List_{NN}}$(i).Delete;
            end
        else if $Singleton(p)==False \wedge Singleton(q)==False$ **then**
            if $d(p,q) < l \times Min(LDI(p), LDI(q)) \wedge GEI(C_{pq})\text{-}GEI(C_{p}) < h$ **then**
                Merge clusters associated with p and q
                $\mathrm{List_{NN}}$(i).Delete;
            end
    end
end

[67]. In NDEC, each remaining singleton in $list_{NN}$ is considered an outlier and based on the problem requirements, any of the above alternatives can be followed to handle the outliers.

4.3.5 NDEC Complexity

Given that n is the size of the dataset and d represents the dimension of the data items, the average time complexity of NDEC is $O(dn \log_2^n)$. The time complexity is calculated as follows: 1) A binary metric tree is constructed for the dataset and is used to identify the nearest neighbours ($O(dn \log_2^n)$), 2) The associations are sorted ($O(n \log_2^n)$), and 3) The list of associations is traversed and clusters are created ($O(n)$).

A strong point of the NDEC clustering algorithm is that it only utilizes neighbourhood distance information. As a result, all pair-wise distances in the given dataset are not required to be calculated. Hence NDEC time complexity is bounded by searching through a metric tree. Since in most cases Euclidean distance is used as the dissimilarity metric, the dimension d is an important factor to be considered when estimating average time complexities. This value can be neglected when we have low-dimensional data.

4.4 Evaluation

The performance of NDEC was assessed using a variety of synthetic and real data sets with a range of data charcateristics and application domains. Table 4.1 shows a description of the artificial and real-world datasets used in this work. Most artificial datasets were obtained from the University of Eastern Finland (UEF) website [90]. In addition to UEF data, three artificial datasets of varying sizes were produced using the *scikit learn* toolkit [91] as suggested in [92] to analyse the effect of dataset size on NDEC parameters. The real-world datasets include data from the UCI Machine Learning Repository [93], the Olivetti Face dataset (ORL), and diffusion tensor imaging (DTI) data from the Johns Hopkins University brain MRI laboratory.

Table 4.1:
Artificial benchmark datasets description.

Artificial					
Name	**Source**	**n**	**d**	**c**	**Description**
Aggregation	**UEF**	788	2	7	Narrow bridges between clusters, uneven-sized clusters [94]
Flame	**UEF**	240	2	2	Arbitrary shaped clusters with constant density [95].
Compound	**UEF**	399	2	6	No clear center in a cluster, densities in the same cluster are various, densities in different clusters are also various [81].
Spiral	**UEF**	312	2	3	Spiral shaped clusters [96]
D31	**UEF**	3100	2	31	Concave shaped clusters [97].
Jain	**UEF**	373	2	2	Arbitrary shaped clusters with non-uniform density [98].
R15	**UEF**	600	2	15	All of the clusters have similar Gaussian distribution [97].
Path_Based	**UEF**	300	2	3	Gaussian distributed clusters surrounded by a circular cluster [96].
Blobs	**scikit-learn**	1K 10K 100K	3	5	Isotropic Gaussian blobs
Moons	**scikit-learn**	1K 10K 100K	2	2	Interleaving half circles
Circle	**scikit-learn**	1K 10K 100K	2	2	A large circle containing a smaller circle
Real					
Name	**Source**	**n**	**d**	**c**	**Description**
Iris	**UCI**	150	4	**3**	Combination of linearly and non-linearly separable classes/clusters
Glass	**UCI**	214	9	6	
Leaf	**UCI**	340	15	30	[99]
Sonar	**UCI**	208	60	2	
ORL	**AT&T**	400	10304	40	Ten different images of each of the 40 subjects taken at different times, with variable lighting, facial details and expressions [100].
DTI	**JHU**	15	60	3	Diffusion tensor imaging data from pediatric subjects aged between 7 and 18 [101]

The Adjusted Rand Index (ARI)[102] and Normalized Mutual Information (NMI)[103] were used as the external evaluation criteria to assess the quality of a clustering solution based on a provided ground truth. ARI indicates the similarity between an obtained clustering and a pre-existing clustering (i.e., a ground truth) and is related to accuracy. ARI also accounts for chance agreements. NMI quantifies the amount of mutual dependence between the two clustering solutions.

To examine the relative efficiency of NDEC compared to other clustering algorithms, four state-of-the-art clustering algorithms were used. As discussed in Section 4.2, clustering algorithms can be categorized into three groups based on their abilities for finding natural clusters within a given data set. The NDEC algorithm is from the ASAD group. For comparison purposes, density peaks clustering (DPC) [84], a well-known non-parametric clustering algorithm in the ASAD group was used. From the ASSD group, DBSCAN [56] and OPTICS, two widespread density-based clustering algorithms were used. From the SSSD group, k-means [55], the most popular partitional clustering algorithm was used.

With respect to the k-means input parameter k, the right number of clusters based on the ground-truth was used. DBSCAN has two parameters, *Eps*, and *Minpts*, where *Eps* determines a threshold on distance range, and *Minpts* determines the minimum number of neighbours required to form a dense region. For DBSCAN, *Eps* was selected from the set of *Eps* values equal to the *Minpts* nearest neighbour distance of each sample, and *Minpts* was selected from the set $\{1, 5, 10, 20\}$.

For OPTICS, *Minpts* was selected from the same set of values used for DBSCAN with the maximum *Eps* value set to the maximum *Minpts* nearest neighbour distance, and steepness parameter ε chosen from the range $0 \leq \varepsilon \leq 1$ [92]. For DPC, as the authors suggested [84], the density was determined by the average distance of 2 percent of the neighbours, and cluster centres were selected manually using the DPC decision graph. In this work, Euclidean distance was used in all cases and clustering parameters were selected based on maximizing both ARI and NMI.

4.4.1 NDEC Parameters Estimation

NDEC has three parameters including number of nearest neighbours (k), distance consistency (l), and entropy consistency (h). Parameter k controls the neighbourhood size. The value of k should be changed incrementally by small steps to prevent unnecessarily big neighbourhood sizes that can decrease the clustering algorithm speed.

Parameter l controls the degree to which samples\sub-clusters will be merged together. Increasing values of l, for the the same k value, means decreasing the number of obtained clusters, and vice versa. Note that, in case of decreasing\increasing k, a corresponding increase\decrease in l is required to have analogous clustering results. Parameter h controls the degree of neighbourhood distance homogeneity within a cluster. The values of parameters k and l are always positive and can be selected just above the value of 1. The estimate of entropy based on sample-spacings can be negative and as a result, the value of h can also be negative.

In the case of k and l we choose $k \in \{3,\ 4,\ ...,\ 11,\ 12\}$ and $l \in \{1.1,\ 1.2,\ 1.3,\ \ldots,\ 2\}$ respectively. The value of h was selected as $h \in \{-1,\ -0.9,\ \ldots,\ 0.9,\ 1\}$. Based on empirical evaluations, the value of h can be fixed at 0.1 for most datasets and the value of k can also be fixed at 4/5 for small datasets. Only the value of l has significant practical effect on clustering. In this section, one potential unsupervised heuristic for selecting appropriate values for these parameters is presented. This heuristic is based on an existing internal clustering validation metric called Density Based Clustering Validation (DBCV)[104]. This index is designed to measure within and between cluster density connectedness and generates values in the range of [-1, 1], with greater values indicating better clustering results [104]. The DBCV index can be used to assess the quality of a clustering solution and is calculated as follows:

Let $O = \{o_1,\ \ldots,\ o_N\}$ be a set of data including N samples with d dimension, *Dist* be an $N \times N$ matrix including pairwise distances $d\,(o_p,\ o_q)$ where $o_p,\ o_q\ \in O$, KNN(o,i) be the distance between sample o and its i^{th} nearest neighbor and $C = \{C_j\}, 1 \leq\ j \leq k$ represent a clustering solution consisting of k clusters for which $\mathrm{N_j}$ is the size of the $\mathrm{j^{th}}$

cluster. Then the following terminologies are used to formulate the DBCV index [104].

1. Core distance of a sample O which belongs to cluster C_j is defined as:

$$c\_dist\,(o) = \left(\frac{\sum_{j=2}^{N_j} \left(\frac{1}{KNN(o,j)} \right)^d}{N_j - 1} \right)^{-\frac{1}{d}} \tag{4.6}$$

2. Mutual reachability distance between two samples $o_i,\ o_j$ is defined as:

$$d_{mreach}\ (o_i,\ o_j) = max\left\{c\_dist\,(o_i)\,,\ c\_dist\,(o_j)\,,\ d\,(o_i,\ o_j)\right\} \tag{4.7}$$

3. Mutual reachability distance graph is a complete graph with all of the samples in O as vertices and edges weighted using the mutual reachability distance between the respective pair of samples.

4. Mutual reachability distance minimum spanning tree (MST_{MRD}) is a minimum spanning tree (MST) of the mutual reachability distance graph created with all samples in O.

5. Density sparseness of a cluster ($DSC(C_j)$) is the maximum edge weight of the internal edges of MST_{MRD}.

6. Density separation between cluster C_i and C_j ($DSPC(C_i,C_j)$ is the minimum reachability distance between internal nodes of their MST_{MRD}s.

7. Validity index of a cluster $V_C(C_j)$ is defined as:

$$V_C\,(C_j) = \frac{\min\left(DSPC\,(C_i,\ C_j) - DSC\,(C_j)\right)}{\max\left(\min\left(DSPC\,(C_i,\ C_j)\right), DSC\,(C_j)\right)} \qquad 1 \leq i \leq k\ ,\ i \neq j \tag{4.8}$$

Table 4.2:
NDEC parameters selected based on maximizing ARI/DBCV

Dataset	Selection Method	k	l	h	ARI	DBCV
Aggregation	ARI	4	1.7	0.1	1	0.3
	DBCV	4	1.7	0.1	1	0.3
Spiral	ARI	4	1.6	0.1	1	0.5
	DBCV	5	1.4	0.1	0.5	0.6
Jain	ARI	5	1.9	0.1	1	0.15
	DBCV	5	1.1	0.1	0.94	0.41
R15	ARI	5	2	0.1	0.99	0.87
	DBCV	5	2	0.1	0.99	0.87
Compound	ARI	4	1.2	0.1	0.99	0.41
	DBCV	4	1.5	0.1	0.94	0.50
Flame	ARI	4	1.5	0.1	0.97	0.64
	DBCV	4	1.5	0.1	0.97	0.64

8. Finally, the validity index of a clustering solution is the weighted average of the validity indexes of all clusters in C and is defined as:

$$DBCV\,(C) = \sum_{\mathrm{j=1}}^{j=\mathrm{k}} \frac{\mathrm{N}_j}{\mathrm{N}}\, V_C\,(C_j) \tag{4.9}$$

Table 4.2 shows NDEC parameters selected based on two different methods for six of the artificial datasets. The first row for each dataset shows parameters selected based on maximizing ARI and its resulting DBCV, whereas in the second row parameters were selected based on maximizing DBCV and its resulting ARI are presented. As the table shows, in most cases, parameters selected using both methods are identical. The correlation between DBCV and ARI has been shown in [67]. It is worth noting that the correlation is positive and consequently an appropriate value for k, l, and h might be selected based on the DBCV measure. However, this correlation is not $+1$, hence selecting parameters based on DBCV might not result in optimal outcomes with respect to ARI.

One helpful feature to investigate regarding NDEC is the effect dataset size n has on the choice of the number of nearest neighbours k, the distance consistency parameter l, and the entropy consistency parameter h. Table 4.3 shows the ARIs obtained after applying NDEC to three artificial datasets with different sizes. For each dataset, the process of generating data remains constant and only the number of drawn samples is changed. From this table, we can observe that the parameters selected for small value of n (1K) and larger values of n (10K, 100K) are identical. These results are specially interesting and they suggest that if a large sufficient sample is available, the parameters k, l, and h may be selected by sampling that dataset.

Table 4.3: Effect of dataset size on NDEC parameters

	Blobs				**Moon**				**Circle**			
Size	**k**	**l**	**h**	**ARI**	**k**	**l**	**h**	**ARI**	**k**	**l**	**h**	**ARI**
1K	10	1.7	0.1	1	10	1.3	0.1	1	10	1.2	0.1	1
10K	10	1.7	0.1	1	10	1.3	0.1	1	10	1.2	0.1	1
100K	10	1.7	0.1	1	10	1.3	0.1	1	10	1.2	0.1	1

4.5 Results and Discussions

In this section, the efficacy of the NDEC clustering algorithm is assessed using several synthetic and real clustering benchmark datasets. Furthermore, the performance of NDEC is compared to that of other state-of-the-art clustering algorithms. In addition, diffusion tensor imaging (DTI) data from the Johns Hopkins University brain MRI laboratory is utilized to demonstrate the usefulness of NDEC a in real-world application.

4.5.1 Benchmark Datasets

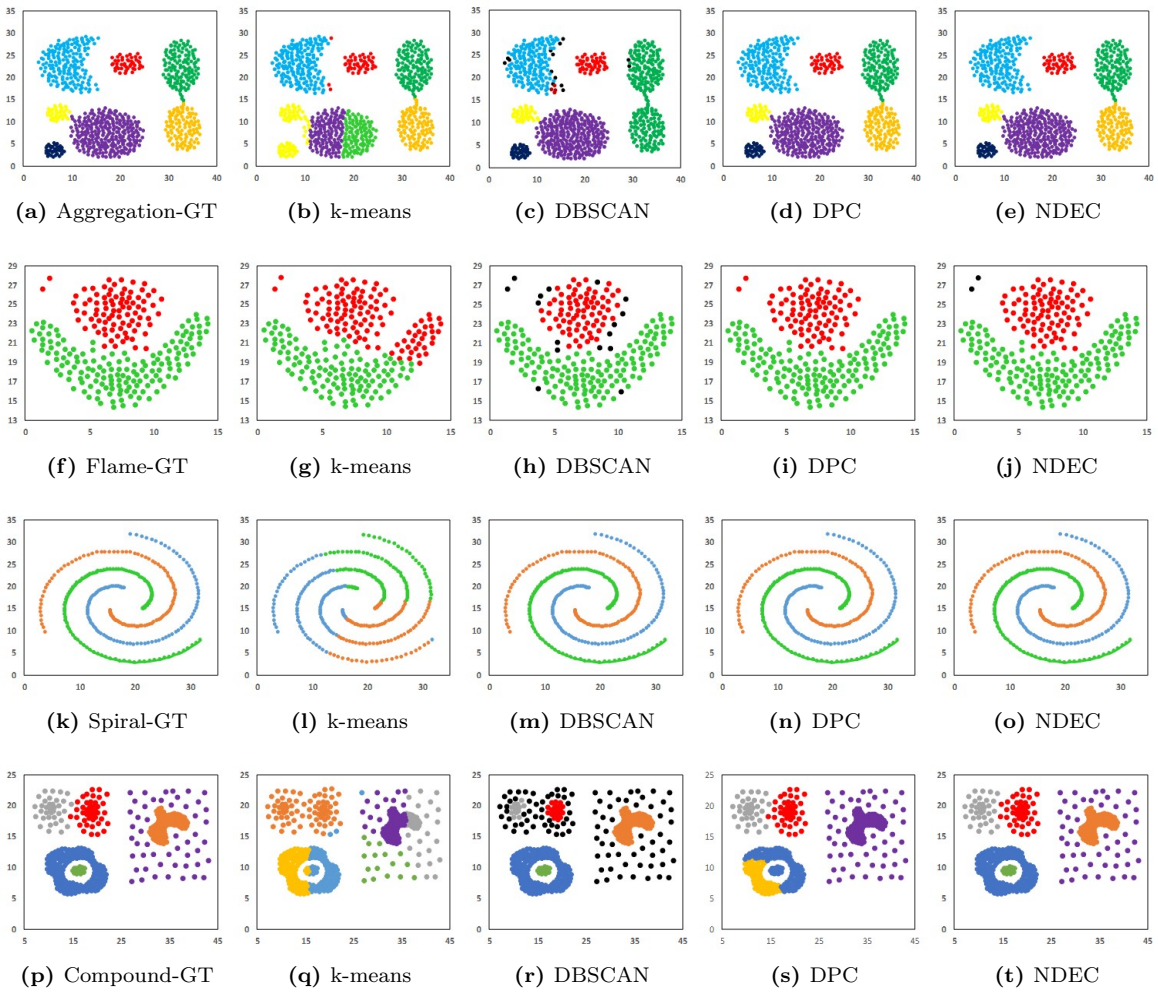
Fig. 4.1 presents the clustering results of k-means, DBSCAN, DPC, and NDEC on four synthetic two-dimensional datasets described in Table 4.1 . Except for Aggregation, most

of the clusters in these datasets do not have clear distinct centres. Hence, k-means has difficulty in identifying the correct clusters. Visual inspection of Fig. 4.1 shows that clusters obtained by NDEC and DPC are the same as the ground truths for Aggregation and Spiral. In contrast, for the Compound dataset NDEC is clearly able to get the closest to the ground truth clusters. In the Compound dataset, the densities within the two upper left corner clusters vary. In addition, the two clusters in the upper right corner have consistent within cluster densities and clearly different densities. Consequently, DBSCAN has difficulty in correctly identifying these clusters. It is worth noting that DPC cannot identify the bigger cluster in the lower left corner since that cluster does not have a distinct centre.

Table 4.4:
A quantitative comparison of NDEC with four state-of-the-art clustering algorithms on benchmark datasets. Note that Aggr and Comp stand for Aggregation and Compound respectively.

	k-means		**DBSCAN**		**OPTICS**		**DPC**		**NDEC**	
Dataset	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
Aggr	0.760	0.878	0.869	0.869	0.993	0.984	**1**	**1**	**1**	**1**
Flame	0.431	0.394	0.857	0.771	0.896	0.805	**1**	**1**	0.971	0.931
Comp	0.536	0.719	0.907	0.780	0.923	0.813	0.592	0.799	**0.997**	**0.994**
Spiral	-0.005	0.000	**1**	**1**	0.307	0.537	**1**	**1**	**1**	**1**
D31	0.952	0.966	0.740	0.882	0.875	0.911	0.934	0.956	**0.994**	**0.990**
Jain	0.304	0.357	0.941	0.862	**1**	**1**	0.643	0.597	**1**	**1**
R15	0.992	0.994	0.916	0.942	0.960	0.969	0.992	**0.994**	**0.996**	0.990
Path	0.461	0.547	0.656	0.704	0.684	0.685	0.530	0.491	**0.934**	**0.886**
Iris	**0.736**	**0.762**	0.568	0.761	0.565	0.745	0.453	0.658	0.675	0.650
Glass	0.247	0.400	0.275	0.515	0.216	0.411	0.092	0.241	**0.384**	**0.429**
Leaf	0.008	0.316	0.187	0.7584	0.228	0.769	0.151	0.623	**0.354**	**0.646**
Sonar	-0.002	0.007	0.000	**0.359**	0.002	0.036	0.019	0.105	**0.193**	0.131

Table 4.4 shows the corresponding ARI and NMI of the clustering solutions obtained using k-means, DBSCAN, OPTICS, DPC, and NDEC on the artificial and the real bench-



(a) Aggregation-GT **(b)** k-means **(c)** DBSCAN **(d)** DPC **(e)** NDEC

(f) Flame-GT **(g)** k-means **(h)** DBSCAN **(i)** DPC **(j)** NDEC

(k) Spiral-GT **(l)** k-means **(m)** DBSCAN **(n)** DPC **(o)** NDEC

(p) Compound-GT **(q)** k-means **(r)** DBSCAN **(s)** DPC **(t)** NDEC

Figure 4.1: Four 2-dimensional benchmark datasets clustering results. Note that the black samples were identified as outliers.

mark datasets described in Table 4.1. As the obtained results show, k-means has difficulty in finding appropriate clusters in most of the analysed datasets. This behaviour can be attributed to the fact that clusters in these datasets do not always have globular shape as k-means assumes. In addition, k-means is susceptible to noise and may not estimate the cluster centres correctly if outliers are present in the dataset under study. The performance of DBSCAN and OPTICS is better than that of k-means in most cases because these algorithms use a density-based definition of a cluster and consequently they are relatively robust to noise and can discover clusters that could not be discovered using k-means.

Fig. 4.2 shows the results of applying NDEC to the Olivetti Faces dataset to categorize the images of the faces of the same woman or man to a cluster. This dataset is a widespread and challenging benchmark used to assess the performance of different unsupervised machine learning algorithms. The Olivetti Faces dataset includes ten different images for each of 40 different persons. Each face is represented by a vector of 10304 features. The clustering task in this dataset is quite challenging since the number of instances is much fewer than the number of features. In this work, the similarity between two images was calculated following the recommendations in [105]. The ARI obtained using NDEC is 0.7 which is significantly higher than the ARI obtained using DPC clustering which is 0.3.

4.5.2 Real-World Applications

In this section, the application of NDEC in neuroscience is evaluated using the JHU-DTI dataset (see Table 4.1). Accurate and robust segmentation of brain white matter (WM) fiber bundles plays a significant role in neuropsychiatric studies and facilitates diagnosing and assessing progression or remission of autism, schizophrenia, and depression. We proposed a WM fiber bundle segmentation method in [58], which involves four main steps including fiber tractography, fiber tracks resampling, similarity matrix calculation, and segmentation using a density-based clustering algorithm. Here, we applied NDEC as a density-based clustering algorithm to reconstructed fibers from difusion tensor imaging tractography. The objective was to segment the data into three interested bundles including

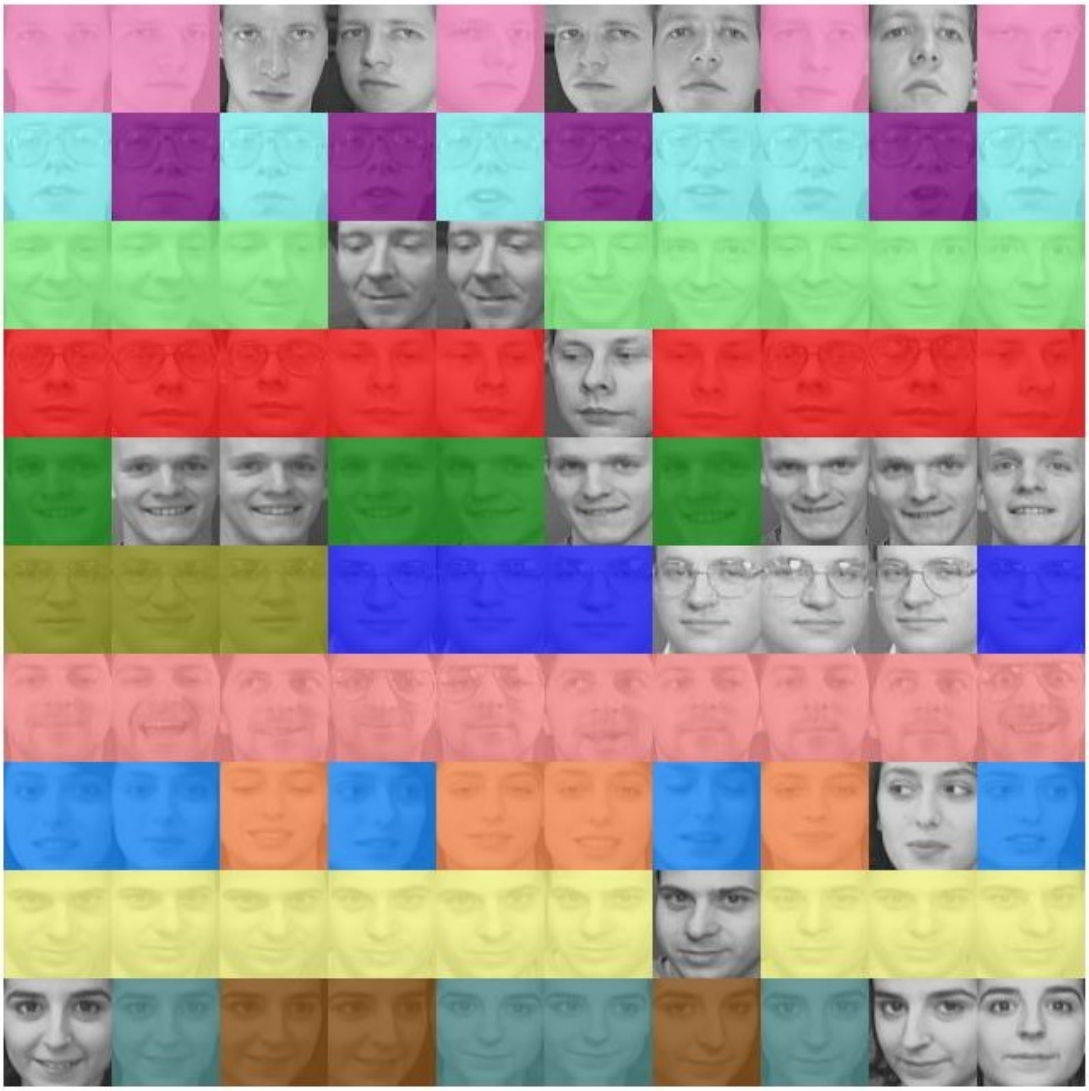


Figure 4.2: Olivetti Faces clustering result obtained by using NDEC. In this figure, the clustering result for the first ten subjects in the dataset is presented. The images of the same color belong to one cluster. The grey images are those that were not assigned to any cluster.

Table 4.5:
WM fiber bundle segmentation results using NDEC and three state-of-the-art clustering algorithms

Dice Ratio (± std)	k-means	DBSCAN	Spectral Clus.	NDEC
IFO	0.75±0.08	0.80±0.05	0.86±0.03	0.91±0.03
ILF	0.77±0.05	0.86±0.04	0.89±0.04	0.94±0.03
Fmajor	0.90±0.06	0.88±0.05	0.89±0.03	0.92±0.02

the inferior fronto-occipital fasciculus (IFO), inferior longitudinal fasciculus (ILF), and the forceps major (Fmajor) [58]. Table 4.5 shows the performance of NDEC compared to three other clustering algorithms with respect to the Dice ratio [106]. In addition to k-means and DBSCAN, spectral clustering [107], a dominant clustering algorithm in the WM segmentation area, was utilized. Spectral clustering is capable of identifying arbitrary-shaped clusters. This algorithm uses eigenvalues of a similarity matrix of the data to perform dimensionality reduction and then applies k-means in a lower dimensional space. As it is shown, NDEC outperforms the other three clustering algorithms with regards to both the mean and standard deviation of the Dice ratios.

Fig. 4.3 presents visualization results for one subject which was selected randomly. As visual inspection shows, the IFO, ILF, and Fmajor results obtained by NDEC are very similar to those obtained by manual segmentation. Furthermore, NDEC is less sensitive to noise compared to the other utilized clustering algorithms. We can observe that k-means, spectral clustering and DBSCAN have trouble in correctly identifying and handling the noise which is pervasive in this domain due to the quality of the tractography results.

4.6 Conclusion

This chapter concentrated on proposing a novel dynamic density based clustering algorithm called Neighbourhood Distance Entropy Consistency (NDEC) and evaluating its absolute and relative performance. NDEC employs both local and global feature space density

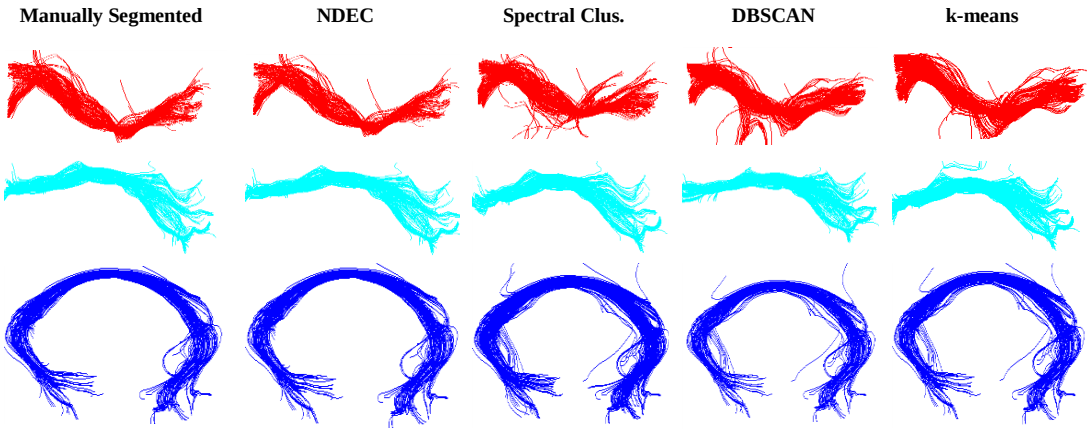


Figure 4.3: Clustering outcomes of one subject selected randomly. The first, second, and third row show IFO, ILF, and Fmajor respectively.

information as well as neighbourhood distance entropy consistency to discover natural clusters existing in data that have arbitrary shapes and densities. The superiority of NDEC over representative algorithms from three different groups of clustering paradigms, with respect to ARI and NMI performance indices, was demonstrated using a variety of benchmark artificial and real clustering datasets. The evaluated clustering paradigms include clustering algorithms which are capable of finding clusters with arbitrary shape and arbitrary density, clusters with arbitrary shape and specific density and clusters with specific shape and specific density.

Chapter 5

Electrophysiological Muscle Classification Using Multiple Instance Learning and Supervised Time Domain Analysis

The work described in this chapter previously appeared in T. Kamali and D. W. Stashuk, A Density-Based Clustering Approach to Motor Unit Potential Characterizations to Support Diagnosis of Neuromuscular Disorders, IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, (7), pp. 956-966, 2017 [1].

5.1 Introduction

In this chapter, a novel electrophysiological muscle classification system is introduced based on the framework proposed in Chapter 3, which focuses on using a dynamic number of classes for characterizing MUPs. To this end, the NDEC clustering algorithm introduced in Chapter 4 is utilized to find representations of several concepts of normality and abnormality in the MUP feature space to be used for MUP characterization. The MUPs sampled from an examined muscle are then characterized, using these increased number of MUP classes, and the MUP characterizations are embedded into a feature vector to represent the examined muscle. The embedded feature vector is then fed to an ensemble of support vector machine (SVM) and nearest neighbor (NN) classifiers to obtain the electrophysiological muscle classification. For 103 sets of MUPs recorded in tibialis anterior muscles, the proposed system had a 97% electrophysiological muscle classification accuracy, which is significantly higher than in previous works.

5.2 Methods

The proposed system has five main steps: 1) MUP representation using morphological and stability features; 2) MUP feature selection using a supervised genetic algorithm; 3) MUP clustering using NDEC; 4) Muscle representation by embedding its MUP characterizations in a feature vector; and 5) Muscle classification using an ensemble of SVM and nearest neighbor classifiers. Details of each step are presented in the following sections.

5.2.1 MUP Representation

Each MUPT is represented by its ensemble of MUPs and an estimated MUP template which is usually calculated by averaging the characteristics of its ensemble of MUPs. This representation assists in analyzing the stability of MUP shape across multiple MU firings.

Broadly speaking, the MUP features can be categorized into two groups: (1) Morphological features, and (2) stability features. Morphological features are extracted from the MUP template and stability features are extracted from its ensemble of MUPs and reflect the MUP morphology changes across all the MUPs in an MUPT. Morphological features, in turn, are classified into three groups with regards to the MUP morphology aspect that they can represent best. These groups include size, shape, and complexity features [108].

Size features are related to the number and sizes of fibers in a given MU. As an example, amplitude and duration are two important size features. The amplitude normally has a value in the range of a few microvolts to several thousand microvolts. The duration can change based on the age of the patient and normally has a value in the range of 5 to 15 ms. Shape features describe the overall shape of a MUP. As an example, thickness is one important shape feature which is measured in milliseconds and represents the width of an MUP. In certain disease states, discriminating normal vs. myopathic and normal vs. neurogenic is tricky. Sometimes MUPs detected from myopathic and neurogenic muscles can have amplitudes comparable to MUPs detected from normal muscles. However, myopathic MUPs usually have a smaller value of thickness due to the loss of muscle fibers and neurogenic MUPs have a larger value of thickness due to reinnervation [10], [4].

Local and global complexity features describe MUP complexity at local and global levels respectively. Normal motor unit fibers are usually spatially dispersed more homogenously compared with fibers in diseased motor units. In addition, the muscle fiber potentials generated by normal motor units usually have less temporal dispersion than those created by diseased motor units. As a result, normal MUPs are more uniform and simple whereas diseased MUPs are usually more complex. As an example, the number of phases is an important global complexity feature which is usually less than four in normal MUPs. In contrast polyphasic MUPs can be detected in neurogenic and myopathic muscles [108]. In this work, each MUP was initially represented by a set of 18 morphological and stability features that are briefly described in Table 5.1. Note that the set of MUPs represented by these features create D_{MUP}. Hence, the MUP training dataset is defined as follows. The i^{th} muscle has a label $Y_i \in \{1,\ 2,\ \ 3\}$ and is represented by the set of MUPs extracted from

Table 5.1:
MUP Morphological and Stability Features

ID	Group	Name	Definition
1	**Size**	Duration	The time difference between the start and end point of an MUP template.
2		Amplitude	The difference in voltage from the minimum positive and maximum negative peak of an MUP template.
3		Area	Rectified MUP template integrated over its duration.
4	**Shape**	Thickness	$Area/Amplitude$
5	**Global Complexity**	Length Index	$\frac{Length-2\times Amplitude}{2\times Amplitude}$, Length is calculated as the summation of the absolute amplitude differences for every 2 consecutive samples within the duration of the MUP template [108].
6		Shape Width	$Area/Length$
7		# of Turns	Number of positive and negative peaks.
8		# of Phases	Discrete number of zero crossings plus one.
9		Fiber Count	Number of near MU fibers [113]
10	**Local Complexity**	Turn Area	$Area/Turns$
11		Phase Area	$Area/Phases$
12		Phase Complexity	$Turns/Phases$
13		Turn Amplitude	$Amplitude/Turns$
14		Turn Length	$Length/Turns$
15		Turn Width	$ShapeWidth/Turns$
16	**Stability**	NF Jiggle	Shape variability of band-pass filtered MUPs
17		Jiggle	Shape variability of raw MUPs recorded using a conventional needle electrode
18		Shimmer Covariance	$\frac{SD\ (MUP_{Dist})}{mean\ (MUP_{Dist})}$

Note that SD stands for the standard deviation. MUP_{Dist}= distances of the MUPs of a MUPT to its MUP template

the clinically-detected EMG signals acquired from the muscle. Let $X_i = \{\overrightarrow{x_{i1}}, \ldots, \overrightarrow{x_{im_i}}\}$ denote the set of m_i MUPs detected in the i^{th} muscle. Each MUP is represented by an M dimensional feature vector $\overrightarrow{x_{ij}} \in R^M$. The MUP dataset is then represented as

$D_{MUP} = \{(X_1, Y_1), \ \ldots, \ (X_N, Y_N)\}$ where N is the number of muscles used to create the dataset.

5.2.2 MUP Feature Selection

MUP features were selected based on maximizing the NDEC-based DBCV index (see Chapter 4 for a detailed description of DBCV) using a genetic algorithm (GA) (see Figure 5.1) with the following setting:

1. The length of each GA chromosome was 21. The first eighteen genes (the same size as the number of features presented in Table 5.1) are binary, where zero elements eliminate the feature and non-zero ones preserve the corresponding feature. The last three genes correspond to the NDEC parameters that were selected based on the intervals shown below:

$$1 < l \leq 7, \ \ 1 < k \leq 3, \ \ 1 < h \leq 2 \tag{5.1}$$

2. The roulette wheel was selected to collect a rich generation. An unbiased one-point crossover was used for generating offspring.

3. The mutation rate was set to 3% to avoid being trapped in phenotypic and genotypic dilemma. The mutation operator for the floating point parameters (i.e., k and h) was defined to pick a new uniform random value between the upper and lower bounds as defined in Eq. 5.1.

4. The merit function was chosen as maximizing within cluster densities while minimizing between cluster densities. This helped to find features that result in better separability among normal and diseased states. This merit function was calculated using DBCV that considers both the density and shape properties of clusters.

5. The population and generation number was set to 50 and 100 respectively.

The GA was utilized to select a subset of the M MUP features and the clustering algorithm parameter values based on their ability to generate a clustering solution with high within cluster densities and low between cluster densities (see Figure 5.1). The output of GA-based MUP feature selection is a new dataset denoted as $D^*_{MUP} = \{(X_1, Y_1), \ \ldots, \ (X_N, Y_N)\}$ where $X_i = \left\{\overrightarrow{x^*_{i1}}, \ldots, \overrightarrow{x^*_{im_i}}\right\}$ and $\overrightarrow{x^*_{ij}} \in R^d$ and $d \leq M$.

5.2.3 MUP Clustering using NDEC

NDEC was applied to a set of MUPs extracted from EMG signals recorded in a set of electrophysiologically similar muscles comprised of healthy/normal muscles and muscles affected, to different extents, by myopathic and neurogenic disorders. The clusters discovered in a MUP feature space were then used as classes for MUP characterization instead of the conventional three classes. Discovering a dynamic number of groups, which represent various possible states of MU health, is challenging due to two main reasons. 1. Neuromuscular disorders are inherently continuous processes and there is no distinct boundary between normal and different stages of disease. Therefore, the clustering algorithm should be able to find clusters with non-convex shape. 2. The level of disease involvement may vary which leads to an unknown number of clusters. Therefore, providing the optimal number of clusters in an initial step of an algorithm is not feasible.

Clusters found in a MUP feature space correspond to different groups of represented MUPs and may be related to different states of MU normality and abnormality. These clusters can be discovered by identifying regions with higher densities isolated from regions with lower densities. The density of a region is determined based on the behavior of the neighborhood distances between sample MUPs; smaller distances relate to higher densities and large distances relate to lower densities. In addition, MUPs belonging to the same cluster should have stable neighborhood distances and consequently stable densities. For this purpose, NDEC was utilized to discover these clusters. After creating $list_{NN}$ that is an abstraction for the MUP feature dataset (see Chapter 4 for a detailed description of the NDEC steps), outlier members were removed using the Modified Thompson Tau rejection

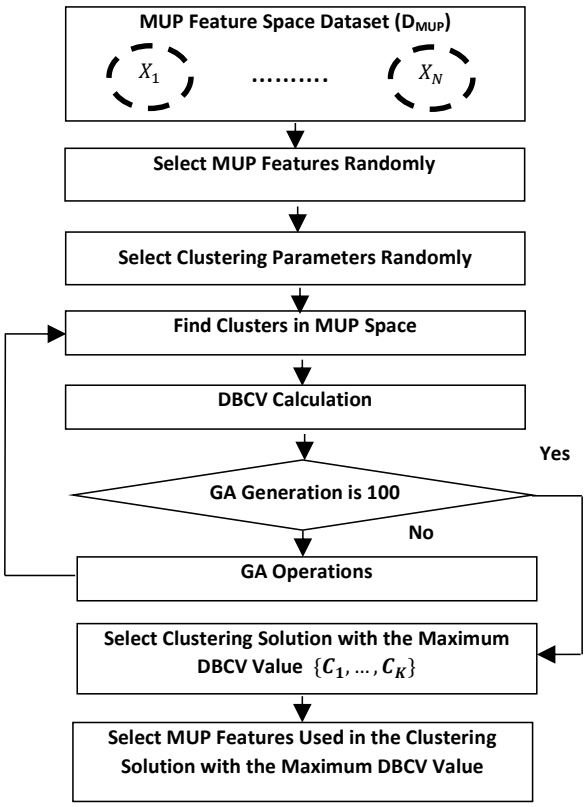


Figure 5.1: Steps for selecting MUP features and finding MUP characterization classes.

rule. For the purpose of finding and removing outliers, the following steps are performed. Let N indicate the length of $list_{NN}$.

1. Calculate the mean ($\overline{d}$) and the standard deviation (S) for all distances (d_i) in $list_{NN}$.

2. Calculate the absolute value of the deviation of each association distance in $list_{NN}$ from ($\bar{d}$):

$$\delta_i = \left| d_i - \bar{d} \right|_{1 \leq i \leq N} \tag{5.2}$$

3. Calculate the modified Thompson Tau (τ) according to the critical value of the

Student's t PDF. In Eq. 5.3, $t_{\alpha/2}$ is the critical Student's t value for =0.05 and df=N-2.

$$\tau = \frac{t_{\alpha/2} \times (n-1)}{\sqrt{n} \times \sqrt{n-2+{t_{\alpha/2}}^2}} \tag{5.3}$$

4. For each distance (d_i) in $list_{NN}$, if $\delta_{\mathrm{i}} > \tau \mathrm{S}$, remove the corresponding association from $list_{NN}$.

NDEC Merge Criteria: The NDEC algorithm used in this work starts merging singletons and creating clusters by traversing $list_{NN}$, each time a tuple *(p,q,d(p,q))* is analyzed, if *d(p,q)* is consistent with both LDI(p) and LDI(q) (Eq. 5.4) and simultaneously LDI(p) and LDI(q) are consistent (Eq. 5.5), then p and q are put into one cluster and their associated tuple is removed from $list_{NN}$.

$$d\left(p,q\right) < l \times \min\left(LDI\left(p\right),\ LDI\left(q\right)\right) \tag{5.4}$$

$$\max(LDI\left(p\right),\ LDI\left(q\right)) < l \times \min\left(LDI\left(p\right),\ LDI\left(q\right)\right) \tag{5.5}$$

If either of p or q belongs to a cluster, the consistency in Eq. 5.4 is checked and instead of LDI in Eq. 5.5, GDI is calculated for non-singleton MUP (i.e. a MUP that is not associated with any cluster) and used in Eq. 5.5 to define the distance consistency criteria and if the distance consistency criteria are met for a singleton MUP, it is assigned to the cluster. If both p and q belong to clusters, the consistency in Eq. 5.4 is checked and GEIs are calculated and used instead of LDI in Eq. 5.5 and instead of l, another user-defined parameter h is used in Eq. 5.5 and if the distance consistency criteria are met, then the two clusters are merged. The merging process is repeated until no change occurs in the length of $list_{NN}$.

5.2.4 Muscle Representation

In this step, a muscle representation is created by embedding its sampled MUPs into a new k-dimensional feature vector $\vec{v} = (v_1, \ldots, v_k)$ where k is the number of discovered clusters (MUP characterization classes) in the MUP feature space. To this end, a matching between the MUPs $\left\{\overrightarrow{x^*_{i1}}, \ldots, \overrightarrow{x^*_{im}}\right\}$ sampled from the muscle X_i and the MUP characterization classes $\{C_j\}_{\forall j \in [1,k]}$ is performed. Consequently, first the degree to which a given MUP belongs to each of the obtained MUP characterization classes ($\{C_j\}_{\forall j \in [1,k]}$) is calculated using Eq. 5.6. Second, the MUP is characterized (labelled) as a member of the characterization class with the best relative fit (Eq. 5.7).

Let p be a given MUP, and q be an arbitrary MUP that belongs to cluster $\mathrm{C_j}$. The degree to which p belongs to $\mathrm{C_j}$ is calculated using Eq. 5.6

$$Bln\,(p, C_j) = min\{d\,(p, q)\}_{\forall q \in C_j} \tag{5.6}$$

The MUP p class label is determined by Eq. 5.7:

$$C_p = \arg min_{C_j}\left\{\frac{Bln\,(p, C_j)}{GDI\ \ (C_j)}\right\}_{\forall j \in [1,k]} \tag{5.7}$$

After MUP characterization labels are determined, the examined muscle is represented in terms of an embedded MUP characterization feature vector based on the labels/characterizations of its sampled MUPs. A muscle X_i is represented by a feature vector $\overrightarrow{M_i}$ of length k as shown in Eq. 5.8, where m_i is the number of MUPs sampled from muscle X_i and x_{ij} is the $\mathrm{j^{th}}$ MUP sampled from muscle X_i.

$$\overrightarrow{M_i} = \left\{\frac{\left|\{x_{ij} \in C_1\}_{1 \le j \le m_i}\right|}{m_i}, \ \ldots, \ \frac{\left|\{x_{ij} \in C_k\}_{1 \le j \le m_i}\right|}{m_i}\right\} \tag{5.8}$$

5.2.5 Electrophysiological Muscle Classification

Finally, as shown in Fig. 5.2, EMC is performed using the embedded MUP characterization feature vector and an ensemble of support vector machines (SVMs) and nearest-neighbour (NN) classifiers. A SVM is selected as a base classifier since it has good generalization capability. The selected SVM has a Gaussian radial basis (RBF) function kernel that is

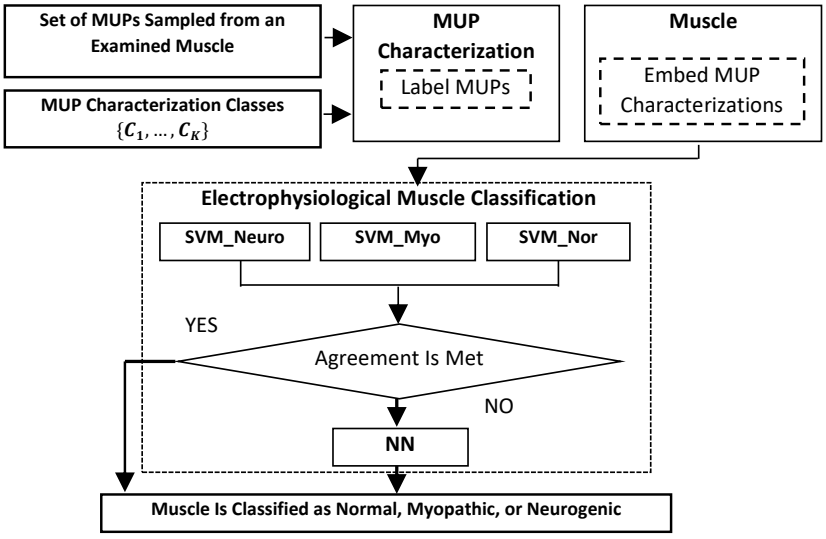


Figure 5.2: Steps for finding the label of an examined muscle. Note that the MUP characterization classes are the output of MUP feature space dataset clustering phase.

expressed as follows:

$$K\left(x, x'\right) = e^{-\gamma||x-x'||^2} \tag{5.9}$$

Where x is the input feature vector to the SVM, $x^{'}$ is the center of the kernel, and γ is the width of the kernel. The use of an ensemble of classifiers can enhance the decision about a pattern to be classified. The EMC problem has three possible class labels while an SVM is a binary classifier; consequently, according to the one-against-all scheme, three base classifiers are considered including: SVM-Nor, SVM-Myo, and SVM-Neuro. SVM-Nor discriminates between normal and diseased muscles, SVM-Myo discriminates between myopathic and

other types of muscles (i.e. normal and neurogenic) and SVM-Neuro discriminates between neurogenic and other types of muscles (i.e. normal and myopathic).

For the purpose of increasing accuracy, the embedded feature vector $\overrightarrow{M_i}$ is multiplied by one of three feature weighting vectors, defined as follows, before being input to the SVM-Myo, SVM-Neuro or SVM-Nor base classifiers, respectively. The basic idea is to assign different weights to different features of the feature space such that the SVM base classifiers learn the decision surface according to the relative importance of feature values in the training dataset.

$$\mathrm{w}_{myo} = \left\{ f_{(C_1,myo)},\ \ldots,\ f_{(C_k,myo)} \right\} \tag{5.10}$$

$$\mathrm{w}_{neuro} = \left\{ f_{(C_1,\mathrm{neuro})},\ \ldots,\ f_{(C_k,\mathrm{neuro})} \right\} \tag{5.11}$$

$$\mathrm{w}_{nor} = \left\{ f_{(C_1,nor)},\ \ldots,\ f_{(C_k,nor)} \right\} \tag{5.12}$$

The elements of the ranking vectors $f_{(C_i,\mathrm{myo})}$, $f_{(C_i,\mathrm{neuro})}$ and $f_{(C_i,\mathrm{nor})}$ are defined using Eq. 5.13, 5.14 and 5.15, respectively, where C_j is the j^{th} MUP characterization class, $n_{(C_j,myo)}$ is the number of times MUPs from C_j are observed in myopathic muscles of the MUP training set D^*_{MUP}, $n_{(C_j,neuro)}$ is the number of times MUPs from C_j are observed in neurogenic muscles of D^*_{MUP} and $n_{(C_j,\mathrm{nor})}$ is the number of times MUPs from C_j are observed in normal muscles of D^*_{MUP}. $\mathrm{N_{myo}}$, $\mathrm{N_{neuro}}$, and $\mathrm{N_{nor}}$ are the number of myopathic, neurogenic and normal muscles that contributed MUPs to D_{MUP^*}, respectively.

$$f_{(C_j,\mathrm{myo})} = \frac{n_{(C_j,myo)}}{N_{myo}} \tag{5.13}$$

$$f_{(C_j,\mathrm{neuro})} = \frac{n_{(C_j,neuro)}}{N_{neuro}} \tag{5.14}$$

$$f_{(C_j,nor)} = \frac{n_{(C_j,\mathrm{nor})}}{N_{nor}} \tag{5.15}$$

If only one of the three base SVM classifiers selects a specific muscle label, the muscle is

classified using this label. As an example, if SVM-Myo selects myopathic, and both SVM-Nor and SVM-Neuro select others, then the muscle is classified as myopathic. Otherwise, due to the complexity of the current decision to be made, a classifier with a complex boundary is required. Note that classifiers with globally complex decision boundaries in general have low generalization capability. As a result, a local classifier should be utilized. As shown in Fig. 5.2, an NN classifier is employed in this work, in cases when there is disagreement among the base SVM classifiers. As an example, if SVM-Myo selects myopathic and SVM-Nor selects normal and SVM-Neuro selects others, then the final decision is made by the NN classifier.

5.3 Evaluation

5.3.1 Data Set

The experiments were performed on EMG data that were sampled from tibialis anterior muscles. The EMG data were acquired under institutional review board (IRB) approval and sanitized of any personal identifying information. The dataset consists of 48 normal muscles with 868 MUPTs, 31 neurogenic muscles with 429 MUPTs and 24 myopathic muscles with 548 MUPTs. The level of disease involvement across the set of studied muscles ranged from slight to moderate to severe. The subjects ranged in age from 21 to 90. The patients with neurogenic disorder had a wide variety of diagnoses including polyneuropathy, polyradiculopathy, and motor neuron disease. The patients with myopathy had inflammatory myopathies or dystrophies such as facioscapulohumeral muscular dystrophy or oculopharyngeal muscular dystrophy.

The data were collected using a concentric needle electrode and a Nicolet Viking EMG machine, with a 10 Hz to 10 kHz bandwidth and a 48 kHz sampling rate at the Mayo Clinic in Phoenix AZ, USA. Needle positioning was performed during low level muscle contraction and then the level of contraction was increased until 40-60 MUPs/s were detected and

then 15 s of EMG signal was acquired. For each muscle, this process was repeated at four spatially distributed locations to get a statistically representative MU sample from the muscle under study.

Each examined muscle was determined to be affected by a myopathic or neurogenic disorder or to be normal by an experienced neurologist based on manual assessments of MUPs detected during low level muscle contraction across all sampled needle positions. Next, MUPTs were extracted from composite EMG signals using decomposition-based quantitative electromyography (DQEMG) [16]. DQEMG is comprised of a set of algorithms for the decomposition of intramuscular EMG signals acquired during isometric contractions. DQEMG decomposes an EMG signal offline by band-pass filtering the signal, detecting the position of MUPs in the filtered signal by a threshold crossing technique, and then grouping the detected MUPs using clustering and knowledge-based classification algorithms.

5.3.2 MUP Clustering Evaluation Criterion

The quality of a clustering solution is evaluated using a relative clustering validity index called Density Based Clustering Validation (DBCV) [104] which considers both density and shape properties of clusters. The DBCV index generates values in the range of [-1, 1], with greater values indicating better clustering solutions (see Chapter 4 for more details).

5.3.3 Electrophysiological Muscle Classification Evaluation

In this work, learning parameters of a SVM, including γ and the penalty factor C [109] were set using grid-search via leave-one out cross validation. To this end, various pairs of (C, γ) were tested and the one with the best cross-validation accuracy was selected. The values for these test pairs were selected following the recommendations in [110]. The performance of the proposed electrophysiological muscle classification system was evaluated using a leave-one-out cross validation method using the embedded MUP characterization feature vectors representing the muscles who contributed MUPs to the MUP training set D^*_{MUP}. Eight

performance indicators were used for this purpose including an accuracy measurement, myopathic, neurogenic and normal sensitivities, and myopathic, neurogenic and normal specificities and the UPA index (Eq. 5.18) which is a combination of the seven mentioned performance indicators [45]. The reported accuracy (A_{Tot}) was calculated by estimating the mean value of individual muscle class (i.e., myopathic, neurogenic, and normal) accuracies. Individual muscle class accuracy was the ratio of the number of correctly classified muscles to the total number of muscles belonging to that class.

$$\overline{Spc} = (Avg\,(Spc_{Myo},\ Spc_{Nro},\ Spc_{Nor}) \times 0.6) \tag{5.16}$$

$$\overline{Sen} = (Avg\,(Sen_{Myo}, Sen_{Nro}, Sen_{Nor}) \times 0.4) \tag{5.17}$$

$$UPA = 0.5\ \times A_{Tot} + 0.5 \times \left(\overline{Spc} + \ \overline{Sen}\right) \tag{5.18}$$

5.3.4 Comparison to State-of-the-art Clustering Algorithms

To examine the relative efficiency of NDEC compared to other clustering algorithms for the task of classifying muscles, four other state of the art clustering algorithms were implemented. These algorithms included k-means from the SSSD group, DBSCAN from the ASSD group, and spectral clustering (SC) [111] and Chameleon from the ASAD group. The electrophysiological muscle classification system based on NDEC is called MC-NDEC and the electrophysiological muscle classification systems based on Chameleon, SC, DBSCAN, and k-means are referred to as MC-Cham, MC-SC, MC-DB, and MC-KM, respectively. Note that all these electrophysiological muscle classification systems are designed based on the method explained in section 5.2.4 and 5.2.5. However, each of these systems, use a different clustering algorithm to define the MUP characterization classes.

5.4 Results and Discussions

In this section, the experimental evaluation of NDEC is first presented and its performance with regard to finding MUP characterization classes, is compared with four state of the art clustering algorithms. Second, the performance of MC-NDEC, is compared with the performance of MC-Cham, MC-SC, MC-DB, and MC-KM, respectively. Finally, the performance of MC-NDEC, is compared with the performance of four previous EMC systems proposed in [37], which are based on three conventional (i.e. normal, myopathic, and neurogenic) classes for MUP characterization.

Finding the MUP characterization classes can be performed in serial or parallel. In serial, a cluster related to MU normality is first found and its members are removed from the dataset (step I) and then additional clusters are discovered using the remaining data (step II). In parallel, all clusters/MUP characterization classes are found simultaneously. Applying NDEC to a MUP feature dataset in both serial and parallel was investigated. Between them, the serial procedure resulted in more accurate outcomes and consequently was selected as the better way to define the MUP characterization classes. The final values selected for NDEC parameters including l, k, and h following step I of the serial procedure are 5, 1.6, and 1.4 respectively. The final values for l, k, and h following step II of the serial procedure are 4, 1.4, and 1.2 respectively.

After applying NDEC to a MUP feature dataset, several clusters are obtained. The cluster with the maximum number of MUPs is assumed to be a representation of MU normality. This selection is performed due to the following facts. First, myopathic and neurogenic muscles usually contain normal MUs. Second, the probability distribution of MUPs related to normal MUs in normal and diseased muscles are the same. In addition, the probability of detecting abnormal MUPs in normal muscles is low. As a result, MUPs representing MU normality are expected to be members of the cluster that has the largest number of MUPs.

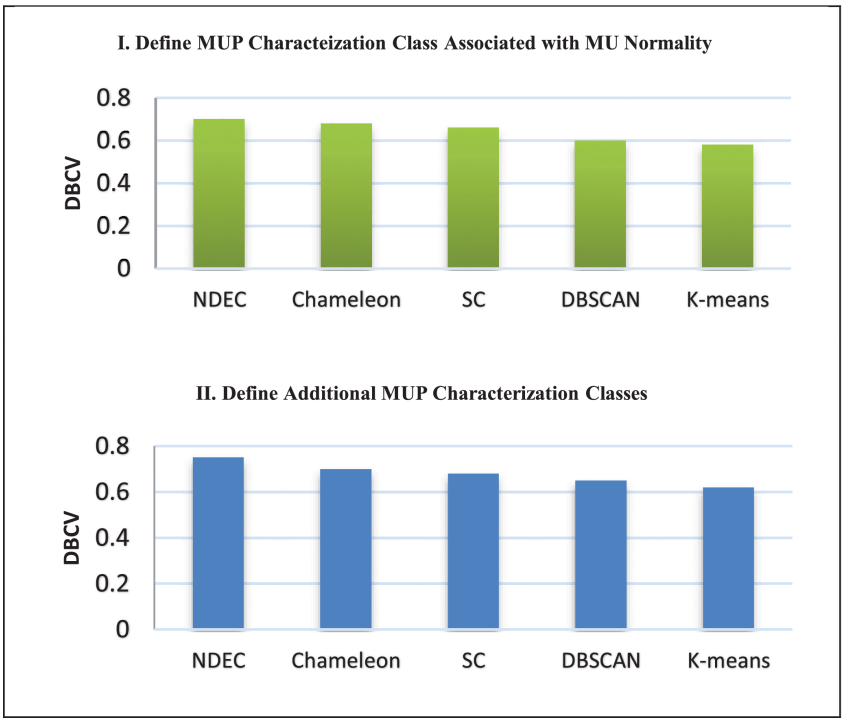


Figure 5.3: Best DBCV values obtained by applying different clustering algorithms to the MUP feature space dataset.

Fig. 5.3 shows the DBCV values of the best clustering solutions obtained using NDEC and the four other clustering algorithms applied in serial to MUP feature datasets. The best clustering solution obtained from applying NDEC, Chameleon, SC, DBSCAN and k-means was composed of 10,8,10,6, and 10 clusters respectively. As can be seen, in both steps of the serial procedure, NDEC resulted in the best clustering solution as measured by the DBCV index. In addition, the ASAD clustering algorithms found the best clustering solutions compared with the ASSD and SSSD groups. The superior clustering solutions obtained using algorithms from the ASAD group supports the notion that clusters found in MUP feature spaces have arbitrary shape and density. Fig. 5.3 also shows that the DBCV values following step II are slightly higher than the DBCV values following step I. This happens because discriminating MUPs related to normal MUs from MUPs related to slightly diseased MUs is more difficult than discriminating MUPs related to MUs affected

by different degrees of myopathic or neurogenic abnormality.

Table 5.2 shows the selected feature sets obtained using the GA and the DBCV index to define MUP characterization classes. Table 5.2 shows that area, shape width and NF jiggle are the most consistent features used to define MUP characterization classes. These features were selected by all five applied clustering algorithms in both steps of the serial procedure. It is worth noting that these selected features are consistent with the qualitative analysis of MUPs performed by physicians. Physicians need to include all aspects (i.e. size, shape, global complexity, local complexity and stability) of MUPs for interpretation. Using a single feature is usually insufficient.

Table 5.2:
Selected Feature Sets for the Purpose Of Defining MUP Characterization Classes

Features Selected to Represent MU Normality					
Clustering Method	**Size**	**Shape**	**Global Complexity**	**Local Complexity**	**Stability**
NDEC	Area	Thickness	Shape Width	Phase Area	NF Jiggle
Chameleon	Area	Thickness	Shape Width	Phase Area	NF Jiggle
SC	Area	Thickness	Shape Width	Phase Area	NF Jiggle
DBSCAN	Area	Thickness	Shape Width	Turn Width	NF Jiggle
K-means	Area	——	Shape Width	Turn Length	NF Jiggle
Features Selected to Define Additional MUP Characterization Classes					
Clustering Method	**Size**	**Shape**	**Global Complexity**	**Local Complexity**	**Stability**
NDEC	Area	Thickness	Shape Width	——	NF Jiggle-Shimmer Covariance
Chameleon	Area	Thickness	Shape Width	——	NF Jiggle-Shimmer Covariance
SC	Area	Thickness	Shape Width	——	NF Jiggle-Shimmer Covariance
DBSCAN	Area	——	Shape Width	Phase Area	NF Jiggle-Shimmer Covariance
K-means	Area	——	Shape Width	Turn Width	NF Jiggle-Shimmer Covariance

Table 5.3:
Performance Indexes of Five EMC Systems

ID	Muscle Classification	Spc_{Myo}	Spc_{Neuro}	Spc_{Normal}	Sen_{Myo}	Sen_{Neuro}	Sen_{Normal}	A_{Tot}	UPA
1	MC-NDEC	98.71	97.18	**100**	**95.65**	**96.66**	**97.91**	**97.02**	**97.44**
2	MC-Cham	94.87	**100**	94.33	91.30	86.66	**97.91**	93.06	93.84
3	MC-SC	98.71	85.91	**100**	86.95	**96.66**	85.41	89.10	90.94
4	MC-DB	**100**	95.77	75.47	52.17	93.33	93.75	84.15	85.14
5	MC-KM	96.15	98.59	69.81	39.13	83.33	**97.91**	80.19	81.24

Table 5.2 also shows that the feature sets used to find a representation of MU normality by most clustering algorithms, are composed of only one feature from the aspect of MUP stability. While when finding the additional MUP characterization classes, two features related to MUP stability were selected. This can be related to the fact that for neurogenic and myopathic MUs, often MUP stability is quite different compared to normal MUs. Consequently, analyzing the MUP stability aspect in abnormal cases is more important than for normal ones.

Fig. 5.4 shows the actual MUP clusters found by applying NDEC. The pie charts represent the percentage of MUPs in each cluster that were recorded in normal, myopathic, and neurogenic muscle, respectively, and show the likelihood that a cluster includes a MUP recorded in a muscle of a specific class. The box plots show the distributions of the MUPs of each cluster with respect to the specific muscle classes. As shown, clusters 2 and 5 predominantly contain MUPs recorded in neurogenic muscles, cluster 10 predominately contains MUPs recorded in myopathic muscle and the rest of the clusters have various mixtures of MUPs recorded from myopathic, normal and neurogenic muscles. These visualization results show that except for cluster 2, a single MUP is not diagnostic and electrophysiological muscle classification must be based on a set of MUPs generated by a representative sample of a muscles MU.

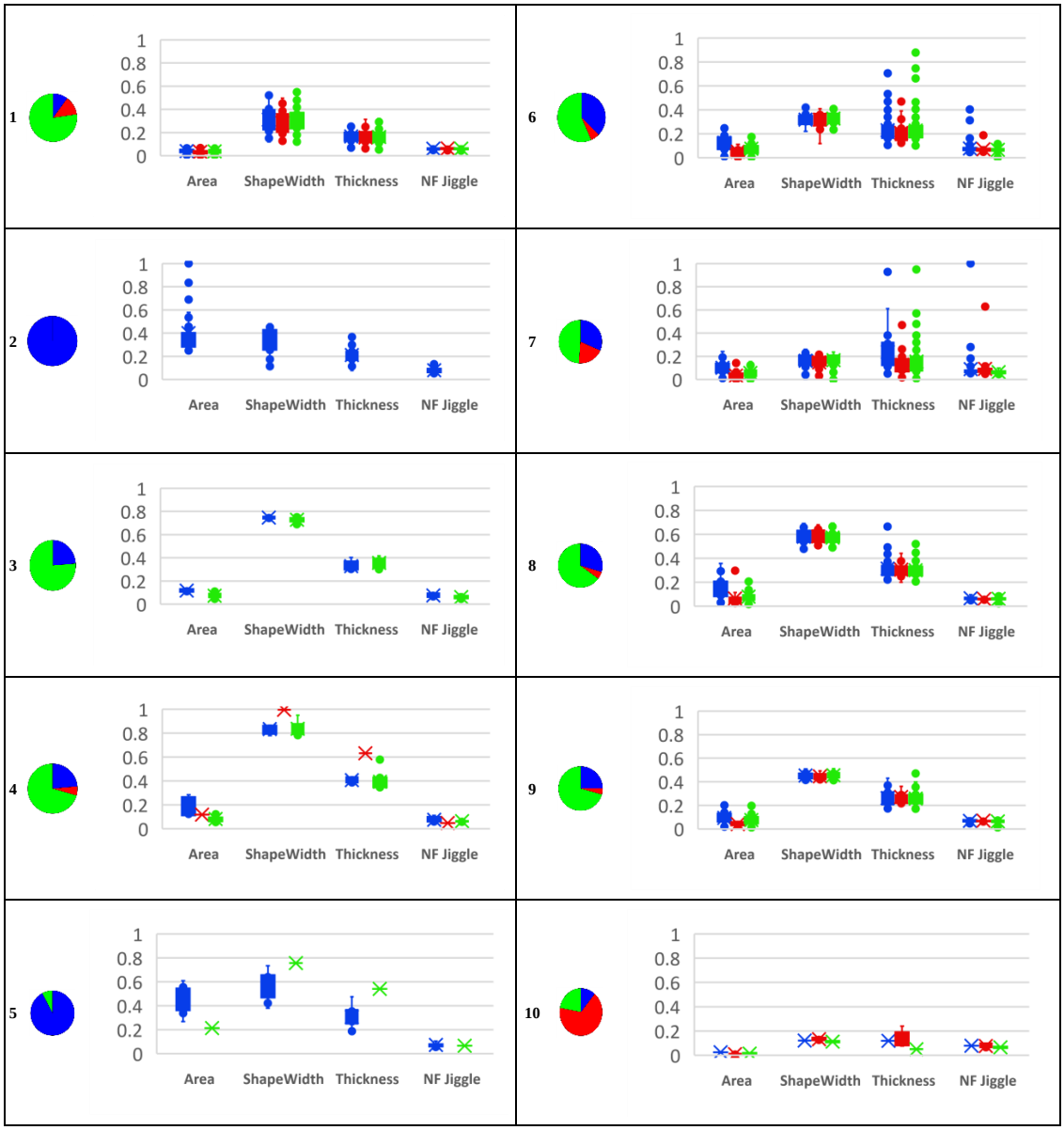


Figure 5.4: Visualization result of clusters obtained by applying NDEC to the MUP feature space dataset. MUPs sampled from normal, myopathic, and neurogenic muscles are represented by green, red, and blue respectively. Cluster 1 was obtained following step I of the serial procedure and clusters 2-10 were obtained following step II of the serial procedure.

Table 5.3 shows the performances indexes of the five implemented EMC systems. The final values selected for learning parameters of SVM including penalty factor C, and kernel width γ are 100, and 0.1 respectively. As the obtained myopathic and neurogenic specificities show, all electrophysiological muscle classification systems are capable of identifying non-myopathic (Spc_{Myo}) and non-neurogenic ($\text{Spc}_{\text{Neuro}}$) muscles with at least 85% accuracy. In addition, with respect to identifying non-myopathic and non-neurogenic muscles, all of the EMC systems evaluated had similar performance. In contrast, EMC techniques using clustering algorithms from the ASAD group resulted in the detection of diseased muscles ($\text{Spc}_{\text{Normal}}$) with at least 94% accuracy whereas methods based on DBSCAN (75%) and k-means (69%) have difficulties identifying disorders.

This shortcoming can be attributed to the fact that clusters in the MUP feature datasets are not always of the same shape, as assumed by k-means, or the same density, as assumed by DBSCAN. In addition, k-means is a type of squared error based clustering algorithm that is highly susceptible to outliers. Comparing acquired specificities leads to the conclusion that discriminating normal muscles from diseased ones is a tough task compared with discriminating non-myopathic and non-neurogenic cases. This happens because transition from normal to diseased is a continuum and there is no distinct boundary between normal and diseased muscles.

Results obtained from sensitivities show that identifying neurogenic and normal muscles is easier compared to identifying myopathic muscles. All of the EMC systems evaluated were able to recognize normal and neurogenic muscles with acceptable sensitivities whereas recognizing myopathic muscles is a difficult task when DBSCAN and k-means clustering is used. Early stage myopathic muscles have very similar characteristics to normal ones, consequently their discrimination requires considering density parameters when clustering the MUP feature datasets, which are only considered by algorithms of the ASAD group.

The obtained accuracies show that using a clustering algorithm from the ASAD group can lead to more accurate electrophysiological muscle classification compared to using clustering methods from the ASSD and SSSD groups, since ASAD algorithms can find

Table 5.4:
Comparison of Different EMC Techniques

ID	**1**	**2**	**3**	**4**	**5**
Method	**MC-EAR-O**	**MC-EAR-U**	**MC-GMM-O**	**MC-GMM-U**	**MC-NDEC**
$\mathbf{A_{Tot}}$	94.1	90.4	92.69	89.66	**97.02**
# of MUP Characterization Classes	3	3	3	3	> 3

natural clusters with arbitrary shape and density in the data. The performance of MC-DB from the ASSD group is better than MC-KM from the SSSD group since DBSCAN first finds a single kernel density estimate for the entire data space and then identifies regions of high density within the data space and consequently can find clusters with arbitrary shape, whereas k-means is only able to find clusters with globular or spherical shapes. Table 5.3 shows that among all implemented electrophysiological muscle classification systems, MC-NDEC is more accurate considering both UPA and total accuracy. In addition, MC-NDEC is more robust and has less variation across the seven performance metrics.

In order to test the significance of the differences among classifier accuracies, an ANOVA test with $\alpha = 0.05$ was performed. The null hypothesis, that all reported mean accuracies are equal, was rejected significantly with a P-value of 0.000983. After rejection of the null hypothesis, the Tukey-Kramer multiple comparison test was performed. The Tukey-Kramer test concluded that there are significant differences among MC-NDEC and the four other muscle classifiers.

Table 5.4 shows a comparison between the accuracy obtained in this work using MC-NDEC and the accuracies reported in previous work [37] that are based on three classes for MUP characterization, using the same tibialis anterior muscle dataset. In this previous work [37], four different approaches for electrophysiological muscle classification were proposed, two of them are based on event association rules (EAR) and the others are based on Gaussian mixture models (GMMs). Multi-class electrophysiological muscle clas-

sification was implemented using both ordered and unordered binarization mappings and resulted in four different groups of muscle classifiers including electrophysiological muscle classification based on GMM using ordered binarization mapping (MC-GMM-O), electrophysiological muscle classification based on GMM using unordered binarization mapping (MC-GMM-U), electrophysiological muscle classification based on EAR using ordered binarization mapping (MC-EAR-O), and electrophysiological muscle classification based on EAR using unordered binarization mapping (MC-EAR-U). For comparison purpose, in Table 5.4, the best results obtained in this previous work [37] for each proposed group of electrophysiological muscle classification technique are presented. As Table 5.4 shows, electrophysiological muscle classification based on NDEC, for which the number of MUP characterization classes is dependent on clustering a MUP feature dataset, outperforms electrophysiological muscle classification systems that are based on only three MUP characterization classes.

5.5 Conclusion

In this chapter, a new EMC system is proposed which classifies muscles based on MUPs detected during isometric contractions. The number of classes for MUP characterization used by the proposed EMC system is dependent on clustering a MUP feature dataset as opposed to conventional EMC systems in which only three classes (i.e. normal, myopathic, and neurogenic) are used for MUP characterization. To this end, a novel dynamic density based clustering algorithm, called Neighbourhood Distance Entropy Consistency (NDEC) is utilized to cluster a reference MUP feature dataset. NDEC uses both local and global MUP feature space density information to discover natural clusters existing in the data that have arbitrary shapes and densities.

The clusters discovered by NDEC are then used as MUP characterization classes. As such, EMC is performed by first characterizing MUPs based on the discovered MUP characterization classes, followed by embedding the MUP characterizations of the MUPs sam-

pled from a muscle to be classified into a feature vector input to an ensemble of SVM and nearest neighbor classifiers to obtain the electrophysiological muscle classification. Results of this work demonstrate that NDEC can be used to discover effective MUP characterization classes. In addition to NDEC, four clustering algorithms using various clustering approaches were implemented to find a relationship between final EMC accuracy and the discovered MUP characterization classes.

The obtained results demonstrate that NDEC can provide superior outcomes with regard to both EMC accuracy and the DBCV relative clustering validation index. The results also show the superior and stable performance of the proposed NDEC-based electrophysiological muscle classification system compared to previously reported electrophysiological muscle classification systems based on only three MUP characterization classes. The proposed clustering algorithm, may also be used as an effective technique in other pattern recognition and medical diagnostic systems in which discovering natural clusters within data is a necessity.

Chapter 6

Electrophysiological Muscle Classification using Unsupervised Time and Spectral Domain Analysis

The work described in this chapter previously appeared in T. Kamali and D. W. Stashuk, Electrophysiological muscle classification using multiple instance learning and unsupervised time and spectral domain analysis, IEEE Transactions on Biomedical Engineering, 2018 [112].

6.1 Introduction

Electrophysiological muscle classification is a crucial step in the diagnosis of neuromuscular disorders. In Chapter 5, an MIL-based EMC system using supervised time domain analysis was presented. In an effort to make the system more robust and accurate so that it can be clinically reliably used, a new MIL-based EMC system is presented in this chapter which is designed based on the framework proposed in Chapter 3.

The evaluation data consists of 63, 83, 93, and 84 sets of MUPs recorded in deltoid, vastus medialis, first dorsal interosseous, and tibialis anterior muscles, respectively. The proposed system discovered representations of MUPs detected in normal, myopathic and neurogenic muscles for each specific muscle type and resulted in an average muscle classification accuracy of 98%, which is higher than in previous works. The results shows that modelling EMC as an instance of MIL solves the traditional problem of characterizing MUPs without full supervision. Furthermore, finding representations of MUP normality and abnormality using morphological, stability, near fiber, and spectral features improves muscle classification accuracy. The proposed method is able to characterize MUPs with respect to disease categories, with no a priori information.

6.2 Methods

This system has five main steps: 1) MUP representation using morphological, stability, and near fiber parameters as well as spectral features extracted from wavelet coefficients; 2) MUP feature selection using unsupervised Laplacian scores; 3) MUP clustering using neighborhood distance entropy consistency to find representations of MUP normality and abnormality; 4) Muscle representation by embedding its MUP cluster associations in a feature vector; and 5) Muscle classification using support vector machines or random forests. The following sections explain the details of each step of the proposed system.

6.2.1 MUP Representation

Time Domain Features

Table 6.1 shows a brief description of the features used to represent an MUPT. Each MUPT can be represented by its ensemble of MUPs and an estimated MUP template which is calculated by ensemble averaging its MUPs. This representation assists in analysing the stability of MUP shapes across multiple MU firings. MUPT time domain features can be categorized into three groups: (1) Morphological features, (2) Stability features, and (3) Near fiber (NF) features. Morphological features are extracted from the MUP template and stability features are extracted from the ensemble of MUPs comprising the MUPT and reflect MUP morphological stability across the MUPs in the MUPT. NF features are extracted from a high-pass filtered MUP template.

High-pass filtering helps to isolate contributions of fibers that are close to the electrode and consequently can potentially provide more robust and detailed information concerning neuromuscular transmission variability. Morphological features, in turn, are classified into three groups with regards to the MUP morphological aspect that they can represent best. These groups include size, shape, and complexity features. Size features are related to the number and sizes of fibers in a given MU. Shape features describe the overall shape of a MUP. Complexity features describe MUP complexity at local and global levels.

Spectral Domain Features

The discrete wavelet transform (DWT), a multi-resolution time-frequency [114] analysis is utilized to represent the relative spectral content of the MUPs. Based on empirical analysis, Daubechies mother wavelet is used because of its high correlation with MUPs. Using DWT results in a high dimensional feature space. Dimensionality is reduced by representing each sub-band by its normalized sub-band energy (NSE) which is defined as the sub-band energy divided by the total energy.

Table 6.1:
MUP Morphological, Stability, and NF Features [1]

ID	Group	Name	Definition
1	**Size**	Duration	The time difference between the start and end point of an MUP template.
2		Amplitude	The difference in voltage from the minimum positive and maximum negative peak of an MUP template.
3		Area	Rectified MUP template integrated over its duration.
4	**Shape**	Thickness	$Area/Amplitude$
5	**Complexity**	Shape Width	$Area/Length$
6		# of Turns	Number of positive and negative peaks.
7		# of Phases	Discrete number of zero crossings plus one.
8	**Stability**	NF Jiggle	Shape variability of NF MUPs
9	**NF**	NF Duration	The time difference between the start and end point of a NF MUP template.
10		NF Dispersion	The time interval between the first and last detected fiber contribution to the NF MUP.
11		NF Count	Number of near MU fibers that are close to the electrode detection surface [113]

6.2.2 MUP Feature Selection

In the approach outlined in this chapter, initially, a MUP is represented by a set of morphological, stability, NF, and DWT features. Usually, some of these features are redundant and/or irrelevant. Hence, an appropriate dimensionality reduction method can enhance the sensitivity and specificity of the developed system. To this end, three main challenges should be considered. 1) MUP training data is unlabelled. 2) For MUP characterization, the local structure of the feature space is more important than the global structure. 3) For the purpose of allowing better diagnosis and treatment planning, the developed EMC system should be interpretable for clinicians, which means that the original meaning of the features needs to be preserved in the new low dimensional subspace.

Reducing dimensionality can be performed by either feature selection or feature extrac-

tion. Feature selection methods select an optimal subset of the original features, whereas the feature extraction methods transform the original features into a new low dimensional subspace, where the new features are a linear or non-linear combination of the original features. Hence, feature selection is used when the original meaning of the features is important.

Feature selection methods can be categorized into four groups including wrapper, embedded, filter, and hybrid methods. A wrapper method selects an optimal subset based on a specified learner. An embedded method selects the best subset during the learning phase of a specific learner. In contrast, the filter method selects a subset with regards to pre-specified evaluation metrics or intrinsic characteristics of the data. The hybrid method combines the advantages of the wrapper and filter based methods and selects a subset with regards to both an independent criterion and a specific learning algorithm.

Considering the MUP dimensionality reduction challenges and the pros and cons of each of the feature selection methods, we used an unsupervised filter feature selection method, called Laplacian score (LS)[115], which is independent of any learning algorithm and can be used to reflect the locality preserving power of each feature. As a result, the features selected in this work, can be used in other EMC systems. Here, for each feature, its LS is computed.

Let L_r denote the LS of the r-th MUP feature, f_{ri} denote the i-th sample of the r-th feature, i = 1, ...,m, and t be a suitable constant. A nearest neighbor graph G with weight matrix S is created based on m nodes. Given the i-th node corresponds to $\mathrm{x_i}$, if either $\mathrm{x_i}$ or $\mathrm{x_j}$ is among the k-nearest neighbors of each other, then i and j are connected and $S_{ij} = e^{-\frac{\left\|x_i - x_j\right\|^2}{t}}$, otherwise $\mathrm{S_{ij}} = 0$. The LS of the r-th feature is computed as follows [115]:

$$f_r = [f_{r1}, ..., f_{rm}]^T, D = diag(S\mathbf{1}), \mathbf{1} = [1, ..., 1]^T, L = D - S \tag{6.1}$$

$$\widetilde{f}_r = f_r - \frac{f_r^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}\mathbf{1}, L_r = \frac{\widetilde{f}_r{}^T L \widetilde{f}_r}{\widetilde{f}_r{}^T D \widetilde{f}_r} \tag{6.2}$$

Smaller LS values correspond to features with greater locality preserving power. Consequently, the MUP features are ranked in ascending order based on their LS values and those which have LS values smaller than a specific threshold are selected as the final MUP features.

6.2.3 MUP Clustering

MUP clustering is performed by utilizing the NDEC clustering algorithm (see Chapter 4). NDEC has four main steps including: 1) Dataset abstraction; 2) Local density estimation using k-nearest neighbours; 3) Generation of sub-clusters based on local and global density consistency; and 4) Generation of final clusters based on entropy consistency. A brief overview of each step is provided below.

Dataset Abstraction

list_{NN} is an abstraction for the dataset D which is defined using Eq. 6.3. Assume p and q are two arbitrary MUPs in D, *d (p, q)* is the symmetric Euclidean distance between them, and $\text{NN}_{\text{k}}(\text{p})$ is a set that comprises of the k (a user defined parameter) nearest neighbours of p.

$$list_{\text{NN}} = \{(p, q, d\,(p, q)\ )|\ p, q\ \in D\ \ \wedge q\ \in NN_k\,(p)\}_{\,\forall p \in D} \tag{6.3}$$

After list_{NN} is created, it is sorted in ascending order with regards to *d (p, q)*.

Local Density Estimation

Local density information (LDI) for MUP p is defined as the average of the distances in the $\text{NN}_{\text{k}}(\text{p})$ set and is calculated for all the MUPs in D.

$$LDI(p) = \frac{1}{k}\sum_{i=1}^{k}\{d_{\text{i}} \in NN_{\text{k}}(p)\} \tag{6.4}$$

Generation of Sub-Clusters

NDEC starts merging singletons and creating clusters by traversing list_{NN}. Each tuple *(p,q,d(p,q))* in the list is analyzed for two conditions. First, the consistency of *d(p,q)* with both *LDI(p)* and *LDI(q)* is investigated. Second, the consistency of *LDI(p)* and *LDI(q)* is analyzed (see Chapter 4). If these conditions are met, p and q are put into one cluster and their associated tuple is removed from list_{NN}. Note that if either of p or q belongs to a cluster, global density information (GDI) is calculated and utilized instead of LDI to define the above distance consistency criteria (see Chapter 4). To define GDI, assume C_p represents the cluster to which MUP p belongs, $\text{list}_{\text{NN}}(\text{C}_\text{p})$ is a list of all associations that belong to C_p, and N_p is the length of $\text{list}_{\text{NN}}(\text{C}_\text{p})$. The GDI for C_p is calculated as the average of the distances (d_j) in $\text{list}_{\text{NN}}(\text{C}_\text{p})$:

$$GDI\left(C_p\right) = \frac{\sum_{\forall d_{j \in list_{NN}\left(C_p\right)}} d_j}{N_p} \tag{6.5}$$

Generation of Final Clusters

While traversing list_{NN} and analyzing an association *(p,q,d(p,q))*, if both p and q belong to a cluster, the global entropy information (GEI) for each cluster is calculated and if the GEIs are consistent (see Chapter 4) the sub-clusters are merged. GEI is defined as follows: First, calculate the order statistics of the distances in $\text{list}_{\text{NN}}(\text{C}_\text{p})$ which are the distances of $\text{list}_{\text{NN}}(\text{C}_\text{p})$ arranged in ascending order $\left(\left\{d^{(1)},\ d^{(2)},\ \ldots,\ d^{(N_p)}\right\}\right)$. Second, estimate the entropy of the distances in $\text{list}_{\text{NN}}(\text{C}_\text{p})$ using Eq. 6.6 based on calculating the m-spacings of the order statistics of the distances in $\text{list}_{\text{NN}}(\text{C}_\text{p})$, where $m = \sqrt{N_p}$. [89].

$$GEI\left(C_p\right) = \frac{1}{N_p - m} \sum_{n=1}^{N_p - m} Log(\frac{N_p + 1}{m}\left(d^{(n+m)} - d^{(n)}\right)) \tag{6.6}$$

The obtained k clusters ($\{C_j\}_{\forall j \in [1,k]}$) will be used as the MUP characterization classes.

6.2.4 Muscle Representation

Similar to the method proposed in section 5.2.4, a muscle is represented by a k-dimensional feature vector $\vec{v} = (v_1, \ldots, v_k)$, where k is the number of obtained clusters in the MUP feature space. To create this embedding, the MUPs $\left\{\overrightarrow{x^*_{i1}}, \ldots, \overrightarrow{x^*_{in}}\right\}$ sampled from the muscle X_i should be characterized/labelled. Let p be a given MUP, and q be an arbitrary MUP that belongs to cluster C_j. The degree to which p belongs to $\mathrm{C_j}$ is calculated using Eq. 6.7:

$$Bln\,(p, C_j) = min\{d\,(p, q)\}_{\forall q \in C_j} \tag{6.7}$$

The MUP p class label is determined using Eq. 6.8:

$$C_p = \arg min_{C_j}\left\{\frac{Bln\,(p, C_j)}{GDI\ \ (C_j)}\right\}_{\forall j \in [1,k]} \tag{6.8}$$

A muscle X_i is then represented by a feature vector $\overrightarrow{M_i}$ using Eq. 6.9 where m_i is the number of MUPs sampled from muscle X_i and $\mathrm{x_{ij}}$ is the $\mathrm{j^{th}}$ MUP sampled from muscle X_i.

$$\overrightarrow{M_i} = \left\{\frac{\left|\{x_{ij} \in C_1\}_{1 \le j \le m_i}\right|}{m_i}, \ \ldots, \ \frac{\left|\{x_{ij} \in C_k\}_{1 \le j \le m_i}\right|}{m_i}\right\} \tag{6.9}$$

6.2.5 Electrophysiological Muscle Classification

To assess the performance of the proposed MIL-based EMC system a global and an ensemble learning model classifier were used. A support vector machine (SVM) [109] was selected as the global model classifier, due to its good generalization capability, and a random forest (RF) [116] was selected as the ensemble model classifier, due to its good transparency and generalization capability.

SVM is a binary classifier whereas the muscle classification problem has three classes, as a result, according to the one-against-all scheme, three classifiers were considered: SVM-

Nor (normal vs. others), SVM-Myo (myopathic vs. others), and SVM-Neuro (neurogenic vs. others). All three SVM classifiers had a Gaussian radial basis (RBF) function kernel (Eq. 6.10), where x is the input feature vector to the SVM, $x^{'}$ is the center of the kernel, and γ is the width of the kernel.

$$K\left(x, x'\right) = e^{-\gamma||x-x'||^2} \tag{6.10}$$

A RF was created from the combination of T decision trees grown from bootstraps sampled from the embedded MUP characterization feature vectors. Individual trees were grown using a greedy procedure, and for each node, Shannon entropy [117] was used as the measure of an impurity criterion.

6.3 Evaluation

The experiments were performed on EMG data that were sampled from four electrophysiologically different groups of muscles (i.e. muscles with different structure and motor control properties). These groups include proximal arm, proximal leg, distal arm, and distal leg muscles. Routine clinical needle EMG was performed in deltoid (DLT), vastus medialis (VM), first dorsal interosseous (FDI), and tibialis anterior (TA) muscles. Table 6.2 provides a detailed description of the number of muscles studied and the corresponding number of MUPTs extracted from each muscle type.

The EMG data were sanitized of any personally identifiable information and approved by the institutional review board (IRB). The individuals participated in this study were between 21 to 90 years of age. The studied muscles had different levels of disease involvement ranging from slight to moderate to severe. A variety of neurogenic diagnoses such as polyradiculopathy, polyneuropathy, and motor neuron disease and myopathic diagnoses such as inflammatory myopathies, and facioscapulohumeral muscular dystrophy were observed in the patients.

Table 6.2:
Evaluation Dataset Description

Muscle			# of Normal		# of Myopathic		# of Neurogenic	
Group		Name	Muscles	MUPTs	Muscles	MUPTs	Muscles	MUPTs
Proximal	Arm	Deltoid	40	915	10	196	13	247
	Leg	VM	60	1230	9	171	13	272
Distal	Arm	FDI	59	1223	8	113	26	426
	Leg	TA	49	1142	10	207	25	431

The data were detected using a Nicolet Viking EMG machine and a concentric needle electrode. The bandwidth was 10 Hz to 10 kHz and the sampling rate was 48 kHz. For each studied muscle, EMG data were collected from four spatially distinct locations to get a statistically representative MU sample. To this end, first, a concentric needle was positioned during low level muscle contraction. Next, the level of contraction was increased until 40-60 MUPs/s were acquired and then 15 s of EMG signal was detected.

Muscles were labelled by an experienced neurologist as normal, myopathic or neurogenic based on manual assessments of MUPs detected during low level muscle contraction across all sampled needle positions. Next, MUPTs were extracted from the composite EMG signals using decomposition-based quantitative electromyography (DQEMG) [16]. The DQEMG algorithms decompose intramuscular EMG signals acquired during isometric contractions. To this end, the signal is band-pass filtered and the position of the MUPs in the filtered signal is detected using a threshold crossing method, and then the detected MUPs are grouped using clustering and knowledge-based classification algorithms. The proposed MIL-EMC system performance was evaluated using seven performance indicators including an accuracy measurement (A_{Tot}), as well as, normal, myopathic and neurogenic sensitivities and specificities.

6.4 Results and Discussions

In this section, the results of the experimental evaluation of the proposed MIL-based EMC system are presented and then its performance is compared with that of four previous EMC systems proposed in [37] which are based on three conventional (i.e. normal, myopathic, and neurogenic) MUP characterization classes.

Fig. 6.1 shows the LS values of both time and spectral domain MUP features of the four different muscle types. In all cases, the number of nearest neighbors used for LS calculation was 4. This number was determined empirically. We constructed a nearest neighbour graph considering 3, 4, 5, and 6 nearest neighbours and among them 4 nearest neighbours resulted in more meaningful clusters with regards to the electrophysiological concepts.

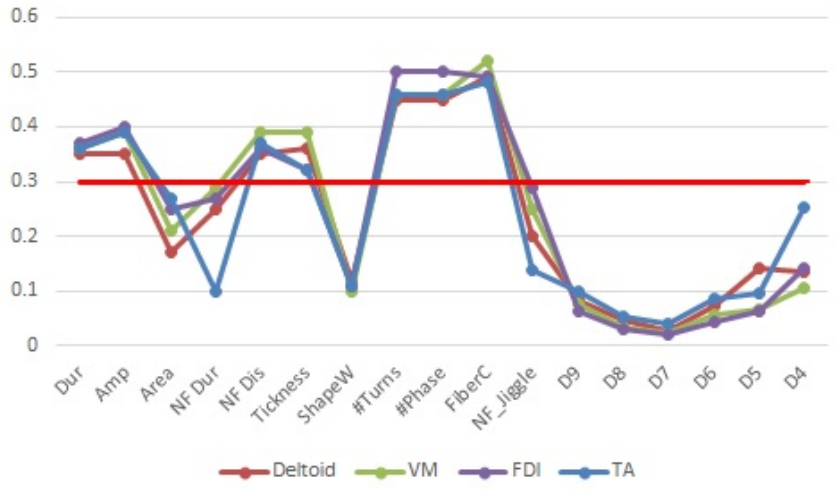


Figure 6.1: LS values of time-domain and DWT MUP features

The LS algorithm ranks MUP features based on their ability to preserve locality, which can be helpful to find representations of MUP normality and abnormality. Hence, the LS cutoff threshold for MUP feature selection should be selected based on the validity and quality of the resulting clustering solution. Note that no ground truth exists for MUP

clustering, as a result, the best clustering solution was determined by its resulting EMC accuracy. Consequently, the final LS cutoff threshold is the one which is associated with the clustering solution that provides the best EMC accuracy. In this work, the LS cutoff threshold was 0.3. In all four muscle classes, Area, ShapeWidth, NF Duration, NF Jiggle, and NSEs of D9, D8, D7, D6, D5, and D4 were selected as the best features.

Table 6.3: MUP dataset clustering results. Note that row sums under Cluster Percentages and column sums under Data Percentages are 100.

			Cluster Percentages			Data Percentages		
Mus.	*Cl#*	*Size*	*%Nor*	*%Myo*	*%Neuro*	*%Nor*	*%Myo*	*%Neuro*
DLT	1	748	92	0	8	**75**	0	25
	2	290	77	7	16	24	11	18
	3	165	0	**100**	0	0	**84**	0
	4	140	0	0	**100**	0	0	**57**
	5	15	33	67	0	1	5	0
VM	1	718	91	1	8	47	3	22
	2	665	87	1	12	**53**	4	30
	3	159	0	**100**	0	0	**93**	0
	4	131	0	0	**100**	0	0	**48**
FDI	1	994	86	0	14	**70**	0	33
	2	420	88	1	11	30	4	11
	3	240	0	0	**100**	0	0	**56**
	4	108	0	**100**	0	0	**96**	0
TA	1	1102	81	3	16	**78**	14	**42**
	2	505	49	31	20	22	**75**	23
	3	151	0	0	**100**	0	0	35
	4	22	0	**100**	0	0	11	0

Table 6.3 shows the results obtained after applying NDEC to the selected MUP feature space of four different muscle types. Depending on the level of disease involvement, myopathic and neurogenic muscles usually have several normal MUs. As such, normality is represented by clusters that have mixed patterns of "normal" MUPs recorded in normal, myopathic and neurogenic muscles. For all four muscle types, NDEC has discovered a

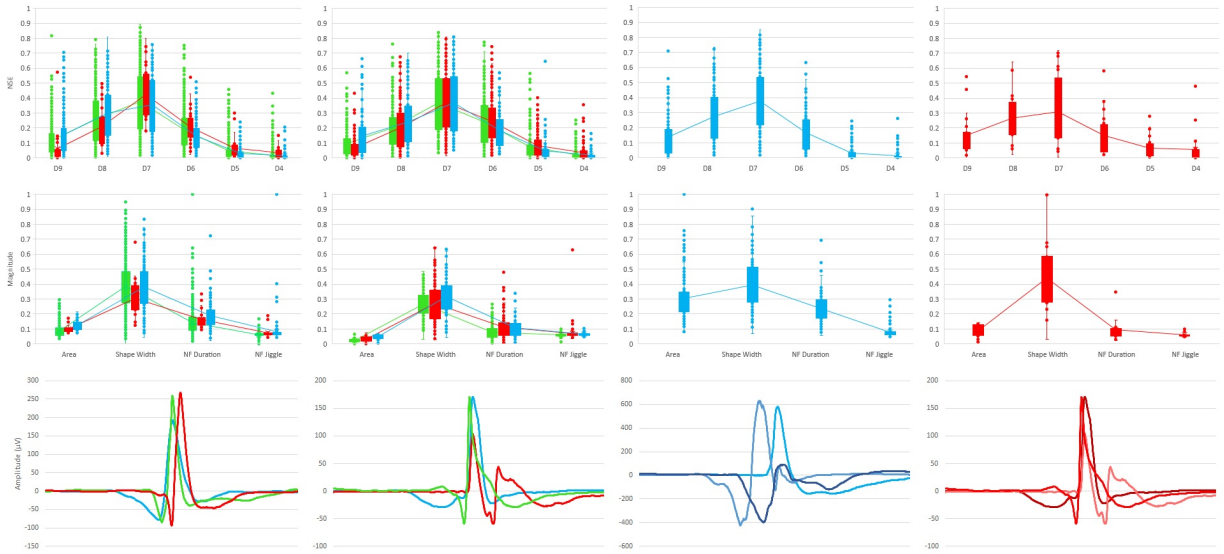


Figure 6.2: TA MUP Clustering Results. For each cluster, 3 representative MUP templates (i.e. MUP templates with close to median cluster area) are presented (25 ms sweep). Green, red and blue represent recordings from normal, myopathic and neurogenic muscles respectively. Time domain features were min-max normalized.

purely myopathic and a purely neurogenic cluster. For the DLT, VM, and FDI musles, the myopathic data percentages for the myopathic clusters (84%, 93%, 96%) are greater than the neurogenic data percentages for the neurogenic clusters (57%, 48%, 56%). Myopathic processes affect muscle fibers based on their spatial distribution and independent of their MU composition, as a result muscle fibers from different MUs can be simultaneously affected. The level of myopathic involvement is dependent on the number of fibers affected. As the disease progresses, most MUs will be affected.

In contrast, neurogenic processes affect motor neurons. All of the muscle fibers belonging to a MU of an affected motor neuron are affected and can become denervated and subsequently reinnervated by a healthy motor neuron. The degree of neurogenic involvement is related to the number of MUs affected. With increased neurogenic involvement the sizes of surviving MUs and the MUPs they generate increase. If reinnnervation is ongoing, unstable MUPs will be recorded. However, a relatively large number of healthy MUs generating normal MUPs may still exist. For the TA muscles, 75% of the MUPs recorded in myopathic muscles belong to cluster 2, which in turn could be interpreted as a

myopathic cluster despite 69% of its MUPs being recorded in non-myopathic muscles. It is worth noting that normal and neurogenic muscles can produce MUPs that look myopathic. This usually happens due to MU sampling phenomenon. Some MUs may have one or two fibers close to the detection surface whereas the bulk of their fibers are far away. In this case, those distant MUPs are going to have similar characteristics as myopathic ones due to dropping amplitude with volume conduction.

The box plots of Fig. 6.2 show the distributions of the features of TA MUPs for each NDEC cluster with respect to the class of muscle they were recorded in (i.e. normal, myopathic, or neurogenic). Cluster 4 contains MUPs recorded in myopathic muscles, cluster 3 contains MUPs recorded in neurogenic muscles and the other two clusters have various mixtures of MUPs recorded in normal, myopathic or neurogenic muscles. The clustering results are consistent with expected disease process effects. A relative increase in high spectral content and low values for size features reflect myopathy whereas, a relative increase in low spectral content and high values for size features reflect neuropathy.

The box plots of Fig. 6.3 show the distributions of the time domain feature values of VM DQEMG data samples for each NDEC cluster with respect to the class of muscle they were recorded in (i.e., normal, myopathic, or neurogenic). As can be seen, each obtained cluster represents a concept. Cluster 4 contains data from MUPTs recorded in neurogenic muscles, cluster 3 contains data from MUPTs recorded in myopathic muscles and the other two clusters have various mixtures of data from MUPTs recorded in normal, myopathic or neurogenic muscles. The clustering results are consistent with expected disease process effects. We can observe that low values for area reflect myopathy whereas, high values for area reflect neuropathy.

Table 6.4 shows the performance of the proposed MIL-based EMC systems. The SVM learning parameters (γ and the penalty factor C) [109] were determined via grid-search using the leave-one out cross validation process. For this purpose, the classification accuracies using various pairs of (C, γ) were calculated and the pair with the best cross-validation accuracy was selected. The number of decision trees, T in the RF was also selected via

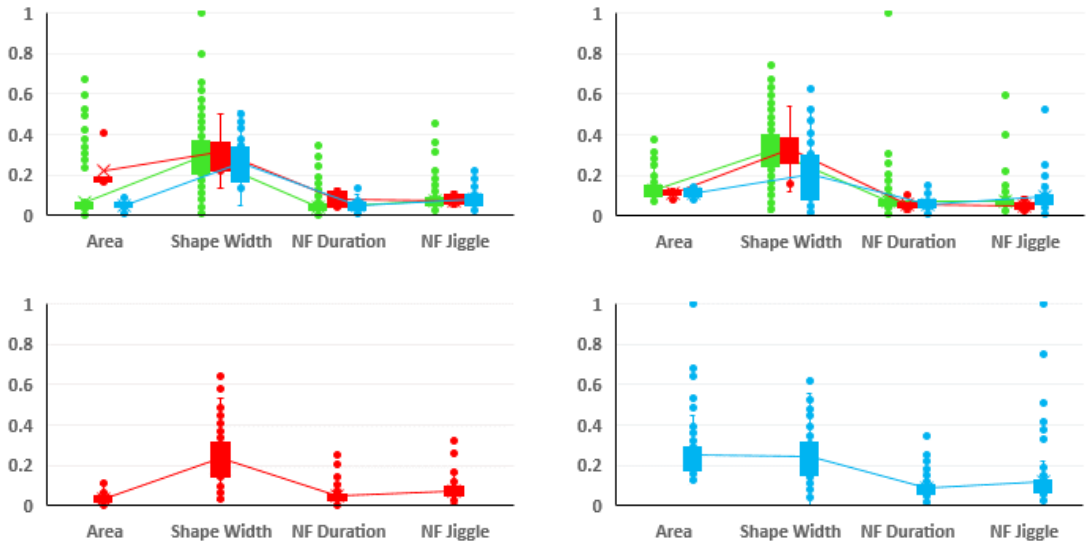


Figure 6.3: Vastus medialis DQEMG clustering results. Data recorded from neurogenic, myopathic, and normal muscles are represnted using blue, red, and green respectively. In order to have better representation, min-max normalization method used for time domain features.

Table 6.4:
Performance Indexes of the proposed MIL-EMC system.

Classifier	Muscle	Spc_{Myo}	Spc_{Neur}	Spc_{Nor}	Sen_{Myo}	Sen_{Neur}	Sen_{Nor}	A_{Tot}
SVM	DLT	100	100	95.65	100	92.30	100	98.41
SVM	VM	100	100	95.45	100	92.31	100	98.78
SVM	FDI	100	100	94.12	87.50	96.15	100	97.85
SVM	TA	98.65	98.31	100	90	96	100	97.61
RF	DLT	100	100	100	100	100	100	100
RF	VM	100	100	95.45	100	92.31	100	98.78
RF	FDI	100	100	94.12	87.50	96.15	100	97.85
RF	TA	97.30	98.31	100	90	92	100	96.42

leave-one out cross validation. The SVMs used were trained using a penalty factor C of 100 and a kernel width γ of 0.1. The number of RF trees used was 68.

As Table 6.4 shows, the normal sensitivities in all muscle types are 100 which shows that the NDEC clusters obtained to represent MUPs recorded in muscles affected by myopathic

and neurogenic disorders are highly diagnostic. In addition, the specificities are often equal to 100 when pathological muscles are considered. This result may be influenced by the small number of myopathic and neurogenic muscles compared with normal ones. However, as the number of myopathic and neurogenic muscles increase, we expect that the clustering algorithm will be better able to find representations of myopathic and neurogenic MUPs. Hence, the final classification accuracies, sensitivities, and specificities should not be altered significantly.

Table 6.5:
Comparison between MIL-EMC and previous EMC techniques

Muscle	OGM-EMC	UGM-EMC	OEA-EMC	UEA-EMC	MIL-EMC
DLT	83.2	74.68	84.87	79.81	100
FDI	86.02	81.5	89.05	84.32	97.85

Table 6.5 shows a comparison between the EMC accuracies obtained in this work and the accuracies reported in a previous effort [37], using the same DLT and FDI muscle datasets. In this previous work, four different EMC methods were implemented including EMC based on a Gaussian mixture model (GMM) using ordered binarization mapping (OGM-EMC), EMC based on a GMM using unordered binarization mapping (UGM-EMC), EMC based on event association rules (EAR) using ordered binarization mapping (OEA-EMC), and EMC based on EAR using unordered binarization mapping (UEA-EMC). All of these methods used only three conventional MUP characterization classes. As Table 6.5 shows, the performance of the proposed MIL-based EMC system is significantly higher than that of the other EMC systems. This improvement can be due to two factors including utilizing spectral and NF features and characterizing MUPs based on a dynamic number of MUP characterization classes which is dependent on clustering a MUP feature dataset.

6.5 Conclusion

Electrophysiological muscle classification can naturally be cast as an instance of a multiple instance learning problem. In this chapter, a new electrophysiological muscle classification system is proposed which transforms the main multiple instance learning problem into a standard supervised learning problem. To this end, morphological, stability, near fiber and spectral distribution features are used to represent MUPs detected during standard clinical EMG examination of a muscle. A mapping function is then utilized to embed the sampled MUPs of the muscle into a single feature vector. Finally, a muscle-level classifier is used to classify the muscle as normal, myopathic or neurogenic.

To determine the mapping function, training sets of suitably represented MUPs were clustered to group MUPs associated with normal and differently diseased MUs. Laplacian scores, unsupervised measures of the locality preserving quality of a feature, were used to select suitable time and spectral domain MUP features. This work considers more than the three conventional groups (i.e. normal, myopathic, and neurogenic) for characterizing MUPs. This improves representation of the effects of disease on both fiber spatial distributions and fiber diameter distributions which lead to a continuity of MUP characteristics. Quantitative results show the superior and stable performance of the proposed MIL-based electrophysiological muscle classification system compared to previous works.

Chapter 7

Conclusions and Future Work

In this chapter, a summary of the thesis contributions and their potential significance are presented. In addition, important research directions that can be pursued for future work are discussed.

7.1 Thesis Contributions

A Novel Density-Based Clustering Algorithm to Discover Natural Clusters

In this thesis, a novel dynamic density based clustering algorithm called Neighbourhood Distance Entropy Consistency (NDEC) is described and its absolute and relative performance when applied to a variety of synthetic and real data sets with a range of data characteristics and application domains is presented and discussed. NDEC employs both local and global feature space density information as well as neighbourhood distance entropy consistency to discover natural clusters existing in data that have arbitrary shapes and densities. NDEC does not require any prior knowledge or assumptions about the number, shape, or density of the clusters. The NDEC clustering algorithm has four main steps including: 1) Local density estimation using the k-nearest neighbours; 2) Generation of

sub-clusters based on local and global density consistency; 3) Generation of final clusters based on neighbourhood distance entropy consistency; and 4) Outlier identification and handling.

Furthermore, one heuristic for selecting NDEC parameters was presented and its validity was further investigated using several artificial datasets with respect to the ARI measure. The superiority of NDEC over representative algorithms from three different groups of clustering paradigms, with respect to the ARI and NMI performance indices, was demonstrated using benchmark artificial and real clustering datasets. The evaluated clustering paradigms include clustering algorithms that are capable of finding clusters with arbitrary shape and arbitrary density, clusters with arbitrary shape and specific density and clusters with specific shape and specific density. In addition, the utility of NDEC was shown in two specific contexts including segmentation of white matter tracts in diffusion tensor imaging and characterizing motor unit potential trains extracted from electromyographic signals. The results show that the NDEC clustering algorithm is helpful for clinical research and practice.

A Novel MIL Framework to Model Electrophysiological Muscle Classification

The objective in MIL-based EMC is to train a classifier f_{MUS}, which can classify previously unseen muscles. This task can be achieved either by following a MUP-level, or muscle-level approach. Using a MUP-level approach, first, a MUP classifier f_{MUS} is trained and subsequently f_{MUS} is defined by combining the outputs of f_{MUP}. Note that in muscle training datasets, labels are only provided at the muscle level. Hence, f_{MUP} must rely on assumptions about the relationship between MUP and muscle labels. Remember that there might be some MUPs recorded from a muscle that do not convey any information about the class label of the muscle, or these MUPs may be even more related to other classes of muscles. Hence, in practice it is not possible to pre-establish a defined relationship between the MUP label and its corresponding muscle label.

Using a muscle-level approach, f_{MUS} is trained directly by defining a supervised rep-

resentation of the muscles. Muscle-level methods, sometimes called "interference pattern analysis", may not provide sufficient sensitivity for clinical application because of superpositions of MUPTs, which makes detection of marginal levels of disease involvement difficult.

While current methods are designed to learn the discriminant information either at the MUP or muscle level, we propose to incorporate both levels of information. To this end, a discriminative embedding of the original feature space is defined based on the characterizations provided by the cluster-adapted MUP classifier. Furthermore, a certain type of information may only be discovered if we consider the discriminative information of the ensemble of MUPs extracted from an EMG signal acquired from the muscle, and not only at the characteristics of the individual MUPs. The proposed method in this thesis incorporates the strengths from both paradigms and hence increases the robustness and accuracy of the obtained results.

A Novel Method to Infer MUP Labels without Full Supervision

Traditional EMC systems assume that MUPs can be classified into just three classes: normal, myopathic, and neurogenic. Additionally, they assume a given relationship between the labels of the MUPs and those of the muscles they were recorded from. Given this assumption, these methods learn a MUP-level classifier. However, the assumption that there are only three classes of MUPs does not necessarily hold in reality. It might well happen that a myopathic muscle is characterized by containing several classes of MUPs. In this work, a novel method was proposed to characterize MUPs into a dynamic number of characterization classes. To this end, the NDEC clustering algorithm was utilized to discover natural clusters existing in a MUP feature space.

This was performed to highlight the effects of disease on both fiber spatial distributions and fiber diameter distributions, which lead to a continuity of MUP characteristics. In addition to NDEC, four clustering algorithms using various clustering approaches were implemented to find a relationship between final muscle classification accuracy and the obtained representations for MU normality and abnormalities. The obtained results demon-

strate that NDEC can provide superior outcomes with regards to both muscle classification accuracy and the DBCV relative clustering validation index.

A Novel Method for Muscle Classification which relies on the Characterization of MUPs

In this thesis, a new muscle classification system was proposed, which classifies muscles based on MUPs detected during isometric contractions. To this end, each muscle is represented by one feature vector that indicates the proportions of classes of MUPs that are present in the EMG signals acquired from that muscle. As such, muscle classification is performed by first characterizing MUPs based on the discovered characterization classes, followed by embedding MUP characterizations into feature vectors input to a standard supervised classifier. SVMs, Random Forests and Nearest Neighbour classifiers were used to classify muscles.

7.2 Future Research

The following interesting challenges can be taken on in the future to extend the methods proposed in this work.

- **Quantify the Diagnostic Information of the Discovered MUP Characterization Classes:** When clustering a MUP feature dataset, several clusters are found. Typically, only a small number of the discovered clusters are meaningful or interesting. Hence, cluster ranking should be applied to select the important clusters. To this end, the importance of a typical cluster should be quantified based on the diagnostic information of that cluster. In this regard, two factors need to be considered including: cluster cohesion, and the probability of observing the cluster in the myopathic, neurogenic, and normal muscle classes.

- **Develop an Appropriate Aggregation Scheme for MUP Characterization Scores:** The notion of a diseased muscle can range from at least one MUP extracted from an EMG signal of the muscle being characterized as diseased to all MUPs being characterized as diseased. Hence, we cannot have any prior knowledge about the exact fraction of diseased MUPs extracted from the EMG signals of a diseased muscle. As a result, a diverse set of aggregation functions should be implemented and analyzed.

- **Develop an Automatic Method to Estimate the NDEC Parameters:** NDEC has three parameters: (1) Number of nearest neighbours (k), (2) Distance consistency (l), and (3) Entropy consistency (h). In this thesis, one potential unsupervised heuristic for selecting appropriate values for these parameters was presented. This heuristic is based on an existing internal clustering validation metric called Density-Based Clustering Validation (DBCV)[104]. It is worth noting that the correlation between DBCV and ARI is positive; consequently an appropriate value for k, l, and h might be selected based on the DBCV measure. However, this correlation is not $+1$, hence selecting parameters based on DBCV might not result in optimal outcomes with respect to ARI. A different direction for future research could be to propose a new heuristic to estimate the NDEC parameters.

- **Develop a Transparent EMC System:** The main role of EMC is guiding the electrophysiological muscle training data analysis and interpretation. This task involves effective and efficient communication of final EMC results to experts. Thus, human-interpretable representation of these results is critical. The MIL-based EMC system proposed in this dissertation has some degree of transparency. The discovered clusters representing MUP characterization classes and the unsupervised filter MUP feature selection method based on Laplacian scores, which reflects the locality preserving power of each feature, are two important examples that help in providing transparency. A transparent muscle-level classifier could also be investigated as an effort to supply experts with sufficient transparency.

References

[1] Tahereh Kamali and Daniel Stashuk. A density-based clustering approach to motor unit potential characterizations to support diagnosis of neuromuscular disorders. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2017.

[2] Roberto Merletti and Philip A Parker. *Electromyography: physiology, engineering, and non-invasive applications*, volume 11. John Wiley & Sons, 2004.

[3] CJ De Luca, RS LeFever, MP McCue, and AP Xenakis. Behaviour of human motor units in different muscles during linearly varying contractions. *The Journal of physiology*, 329(1):113–128, 1982.

[4] Yunfen Wu, María Ángeles Martínez Martínez, and Pedro Orizaola Balaguer. Overview of the application of emg recording in the diagnosis and approach of neurological disorders. In *Electrodiagnosis in New Frontiers of Clinical Research*. InTech, 2013.

[5] Charles Farkas, Andrew Hamilton-Wright, Hossein Parsaei, and Daniel W Stashuk. A review of clinical quantitative electromyography. *Critical Reviews in Biomedical Engineering*, 38(5), 2010.

[6] Erik Stlberg and Björn Falck. The role of electromyography in neurology. *Electroencephalography and clinical Neurophysiology*, 103(6):579–598, 1997.

[7] Erik V Stålberg and Masahiro Sonoo. Assessment of variability in the shape of the motor unit action potential, the jiggle, at consecutive discharges. *Muscle & nerve*, 17(10):1135–1144, 1994.

[8] Anders Fuglsang-Frederiksen. The role of different emg methods in evaluating myopathy. *Clinical neurophysiology*, 117(6):1173–1189, 2006.

[9] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[10] Walter R Frontera and Julien Ochala. Skeletal muscle: a brief review of structure and function. *Calcified tissue international*, 96(3):183–195, 2015.

[11] Richard Kendall and Robert A Werner. Interrater reliability of the needle examination in lumbosacral radiculopathy. *Muscle & nerve*, 34(2):238–241, 2006.

[12] Zoia C Lateva and Kevin C McGill. Estimating motor-unit architectural properties by analyzing motor-unit action potential morphology. *Clinical neurophysiology*, 112(1):127–135, 2001.

[13] Mathias Tröger and Reinhard Dengler. The role of electromyography (emg) in the diagnosis of als. *Amyotrophic lateral sclerosis and other motor neuron disorders: official publication of the World Federation of Neurology, Research Group on Motor Neuron Diseases*, 1:S33–40, 2000.

[14] Dan Stashuk. Emg signal decomposition: how can it be accomplished and used? *Journal of Electromyography and Kinesiology*, 11(3):151–173, 2001.

[15] Ronald S LeFever and Carlo J De Luca. A procedure for decomposing the myoelectric signal into its constituent action potentials-part i: technique, theory, and implementation. *IEEE transactions on biomedical engineering*, (3):149–157, 1982.

[16] Daniel William Stashuk. Decomposition and quantitative analysis of clinical electromyographic signals. *Medical engineering & physics*, 21(6):389–404, 1999.

[17] Christodoulos I Christodoulou and Constantinos S Pattichis. Unsupervised pattern recognition for the classification of emg signals. *IEEE Transactions on Biomedical Engineering*, 46(2):169–178, 1999.

[18] Daniel Zennaro, Peter Wellig, Volker M Koch, George S Moschytz, and Thomas Laubli. A software package for the decomposition of long-term multichannel emg signals using wavelet coefficients. *IEEE Transactions on Biomedical Engineering*, 50(1):58–69, 2003.

[19] Christos D Katsis, Yorgos Goletsis, Aristidis Likas, Dimitrios I Fotiadis, and Ioannis Sarmas. A novel method for automated emg decomposition and muap classification. *Artificial Intelligence in Medicine*, 37(1):55–64, 2006.

[20] Christos D Katsis, Themis P Exarchos, Costas Papaloukas, Yorgos Goletsis, Dimitrios I Fotiadis, and Ioannis Sarmas. A two-stage method for muap classification based on emg decomposition. *Computers in Biology and Medicine*, 37(9):1232–1240, 2007.

[21] Miki Nikolic and Christian Krarup. Emgtools, an adaptive and versatile tool for detailed emg analysis. *IEEE Transactions on Biomedical Engineering*, 58(10):2707–2718, 2011.

[22] DANIEL Stashuk and HUBERT De Bruin. Automatic decomposition of selective needle-detected myoelectric signals. *IEEE transactions on biomedical engineering*, 35(1):1–10, 1988.

[23] Daniel W Stashuk and RK Naphan. Probabilistic inference-based classification applied to myoelectric signal decomposition. *IEEE transactions on biomedical engineering*, 39(4):346–355, 1992.

[24] MH Hassoun, Chuanming Wang, and AR Spitzer. Nnerve: neural network extraction of repetitive vectors for electromyography. ii. performance analysis. *IEEE transactions on biomedical engineering*, 41(11):1053–1061, 1994.

[25] Sarbast Rasheed, Daniel W Stashuk, and Mohamed S Kamel. Integrating heterogeneous classifier ensembles for emg signal decomposition based on classifier agreement. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):866–882, 2010.

[26] Abdulhamit Subasi. Classification of emg signals using combined features and soft computing techniques. *Applied soft computing*, 12(8):2188–2198, 2012.

[27] EW Abel, PC Zacharia, A Forster, and TL Farrow. Neural network analysis of the emg interference pattern. *Medical engineering & physics*, 18(1):12–17, 1996.

[28] Christodoulos I Christodoulou and Constantinos S Pattichis. Combining neural classifiers in emg diagnosis. In *6th European Congress on Intelligent Techniques and Soft Computing. EUFIT98*, volume 3, pages 1837–41, 1998.

[29] Tameem M Adel, Benn E Smith, and Daniel W Stashuk. Muscle categorization using pdf estimation and naive bayes classification. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2619–2622. IEEE, 2012.

[30] Nihal Fatma Güler and Sabri Koçer. Classification of emg signals using pca and fft. *Journal of Medical Systems*, 29(3):241–250, 2005.

[31] Prodromos A Kaplanis, Constantinos S Pattichis, Damjan Zazula, et al. Multiscale entropy-based approach to automated surface emg classification of neuromuscular disorders. *Medical & biological engineering & computing*, 48(8):773–781, 2010.

[32] Abdulhamit Subasi. Medical decision support system for diagnosis of neuromuscular disorders using dwt and fuzzy support vector machines. *Computers in Biology and Medicine*, 42(8):806–815, 2012.

[33] Abdulhamit Subasi. Classification of emg signals using pso optimized svm for diagnosis of neuromuscular disorders. *Computers in biology and medicine*, 43(5):576–586, 2013.

[34] Keleş Selami. Classification of emg signals using decision tree methods. 2012.

[35] Ercan Gokgoz and Abdulhamit Subasi. Comparison of decision tree algorithms for emg signal classification using dwt. *Biomedical Signal Processing and Control*, 18:138–144, 2015.

[36] ABM Sayeed Ud Doulah, Shaikh Anowarul Fattah, Wei-Ping Zhu, M Omair Ahmad, et al. Wavelet domain feature extraction scheme based on dominant motor unit action potential of emg signal for neuromuscular disease classification. *IEEE Trans. Biomed. Circuits and Systems*, 8(2):155–164, 2014.

[37] Meena AbdelMaseeh, Tsu-Wei Chen, Pascal Poupart, Benn Smith, and Daniel Stashuk. Transparent muscle characterization using quantitative electromyography: Different binarization mappings. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(3):511–521, 2014.

[38] Lou J Pino and Daniel W Stashuk. Using motor unit potential characterizations to estimate neuromuscular disorder level of involvement. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 4138–4141. IEEE, 2008.

[39] Sarbast Rasheed, Daniel Stashuk, and Mohamed Kamel. Adaptive fuzzy k-nn classifier for emg signal decomposition. *Medical engineering & physics*, 28(7):694–709, 2006.

[40] Gurmanik Kaur, Ajat Shatru Arora, and VK Jain. Multi-class support vector machine classifier in emg diagnosis. *WSEAS Transactions on Signal Processing*, 5(12):379–389, 2009.

[41] GURMANIK Kaur, AS Arora, and VK Jain. Emg diagnosis via ar modeling and binary support vector machine classification. *Int J Eng Sci Technol*, 2(6):1767–1772, 2010.

[42] Andrzej P Dobrowolski, Mariusz Wierzbowski, and Kazimierz Tomczykiewicz. Multiresolution muaps decomposition and svm-based analysis in the classification of neuromuscular disorders. *Computer methods and programs in biomedicine*, 107(3):393–403, 2012.

[43] Kazimierz Tomczykiewicz, Andrzej P Dobrowolski, and Mariusz Wierzbowski. Evaluation of motor unit potential wavelet analysis in the electrodiagnosis of neuromuscular disorders. *Muscle & nerve*, 46(1):63–69, 2012.

[44] T Kamali, R Boostani, and H Parsaei. A hybrid classifier for characterizing motor unit action potentials in diagnosing neuromuscular disorders. *Journal of Biomedical Physics & Engineering*, 3(4):145, 2013.

[45] Tahereh Kamali, Reza Boostani, and Hossein Parsaei. A multi-classifier approach to muap classification for diagnosis of neuromuscular disorders. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):191–200, 2014.

[46] Lou Joseph Pino. Neuromuscular clinical decision support using motor unit potentials characterized by'pattern discovery'. 2009.

[47] Jaume Amores. Vocabulary-based approaches for multiple-instance data: a comparative study. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4246–4250. IEEE, 2010.

[48] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010.

[49] Hossein Hajimirsadeghi and Greg Mori. Multi-instance classification by max-margin training of cardinality-based markov networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1839–1852, 2017.

[50] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71, 1997.

[51] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003.

[52] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998.

[53] Jérôme Louradour and Hugo Larochelle. Classification of sets using restricted boltzmann machines. *arXiv preprint arXiv:1103.4896*, 2011.

[54] Lin Dong. *A comparison of multi-instance learning algorithms*. PhD thesis, The University of Waikato, 2006.

[55] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[56] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[57] George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[58] Tahereh Kamali and Daniel Stashuk. Automated segmentation of white matter fiber bundles using diffusion tensor imaging data and a new density based clustering algorithm. *Artificial intelligence in medicine*, 73:14–22, 2016.

[59] BS Everitt, S Landau, M Leese, and D Stahl. Cluster analysis: Wiley series in probability and statistics, 2011.

[60] Minh-Ha Nguyen and Sameer Alam. Airspace collision risk hot-spot identification using clustering models. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):48–57, 2018.

[61] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on knowledge and data engineering*, 16(11):1370–1386, 2004.

[62] Annemie Ribbens, Jeroen Hermans, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Unsupervised segmentation, clustering, and groupwise registration of heterogeneous populations of brain mr images. *IEEE transactions on medical imaging*, 33(2):201–224, 2014.

[63] Liang-Jie Zhang, Shuxing Cheng, Carl K Chang, and Qun Zhou. A pattern-recognition-based algorithm and case study for clustering and selecting business services. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(1):102–114, 2012.

[64] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[65] Mei Chen, Longjie Li, Bo Wang, Jianjun Cheng, Lina Pan, and Xiaoyun Chen. Effectively clustering by finding density backbone based-on knn. *Pattern Recognition*, 60:486–498, 2016.

[66] Juan Lu, Zhiguo Gong, and Xuemin Lin. A novel and fast simrank algorithm. *IEEE transactions on knowledge and data engineering*, 29(3):572–585, 2017.

[67] Davoud Moulavi. Finding, evaluating and exploring clustering alternatives unsupervised and semi-supervised. 2014.

[68] Alessandro Lulli, Matteo Dell'Amico, Pietro Michiardi, and Laura Ricci. Ng-dbscan: scalable density-based clustering for arbitrary data. *Proceedings of the VLDB Endowment*, 10(3):157–168, 2016.

[69] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[70] Raymond T. Ng and Jiawei Han. Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, 14(5):1003–1016, 2002.

[71] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. In *ACM Sigmod Record*, volume 25, pages 103–114. ACM, 1996.

[72] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[73] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: an efficient clustering algorithm for large databases. In *ACM Sigmod Record*, volume 27, pages 73–84. ACM, 1998.

[74] Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *VLDB*, volume 98, pages 428–439, 1998.

[75] Alexander Hinneburg, Daniel A Keim, et al. An efficient approach to clustering in large multimedia databases with noise. In *KDD*, volume 98, pages 58–65, 1998.

[76] Levent Ertöz, Michael Steinbach, and Vipin Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 47–58. SIAM, 2003.

[77] Bao-Zhi Qiu, Xi-zhi Zhang, and Jun-yi Shen. Grid-based clustering algorithm for multi-density. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 3, pages 1509–1512. IEEE, 2005.

[78] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, volume 28, pages 49–60. ACM, 1999.

[79] Sheng Li, Lusi Li, Jun Yan, and Haibo He. Sde: A novel clustering framework based on sparsity-density entropy. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[80] Soumaya Louhichi, Mariem Gzara, and Hanêne Ben-Abdallah. Unsupervised varied density based clustering algorithm using spline. *Pattern Recognition Letters*, 93:48–57, 2017.

[81] Charles T Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on computers*, 100(1):68–86, 1971.

[82] Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

[83] Noha A Yousri, Mohamed S Kamel, and Mohamed A Ismail. A distance-relatedness dynamic model for clustering high dimensional data of arbitrary shapes and densities. *Pattern Recognition*, 42(7):1193–1209, 2009.

[84] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[85] Matthew Browne. Regularized tessellation density estimation with bootstrap aggregation and complexity penalization. *Pattern Recognition*, 45(4):1531–1539, 2012.

[86] Noha A Yousri, Mohammed A Ismail, et al. Adaptive similarity search in metric trees. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 419–424. IEEE, 2007.

[87] Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

[88] Oldrich Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 54–59, 1976.

[89] Erik G Miller. A new class of entropy estimators for multi-dimensional densities. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 3, pages III–297. IEEE, 2003.

[90] Pasi Fränti et al. Clustering basic benchmark, 2015.

[91] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[92] Avory Christopher Bryant and Krzysztof J Cios. Rnn-dbscan: A density-based clustering algorithm using reverse nearest neighbor density estimates. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[93] Moshe Lichman et al. Uci machine learning repository, 2013.

[94] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):4, 2007.

[95] Limin Fu and Enzo Medico. Flame, a novel fuzzy clustering method for the analysis of dna microarray data. *BMC bioinformatics*, 8(1):3, 2007.

[96] Hong Chang and Dit-Yan Yeung. Robust path-based spectral clustering. *Pattern Recognition*, 41(1):191–203, 2008.

[97] Cor J. Veenman, Marcel J. T. Reinders, and Eric Backer. A maximum variance cluster algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 24(9):1273–1280, 2002.

[98] Anil K Jain and Martin HC Law. Data clustering: A users dilemma. In *International conference on pattern recognition and machine intelligence*, pages 1–10. Springer, 2005.

[99] Pedro FB Silva, Andre RS Marcal, and Rubim M Almeida da Silva. Evaluation of features for leaf discrimination. In *International Conference Image Analysis and Recognition*, pages 197–204. Springer, 2013.

[100] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142. IEEE, 1994.

[101] Kenichi Oishi, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Jiangyang Zhang, John T Hsu, Michael I Miller, Peter CM van Zijl, Marilyn Albert, et al. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. *Neuroimage*, 46(2):486–499, 2009.

[102] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[103] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[104] Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 839–847. SIAM, 2014.

[105] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009.

[106] Miguel Murguía and José Luis Villaseñor. Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classifications. In *Annales Botanici Fennici*, pages 415–421. JSTOR, 2003.

[107] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[108] Meena Abdelmaseeh, Benn Smith, and Daniel Stashuk. Feature selection for motor unit potential train characterization. *Muscle & nerve*, 49(5):680–690, 2014.

[109] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[110] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[111] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[112] Tahereh Kamali and Daniel W Stashuk. Electrophysiological muscle classification using multiple instance learning and unsupervised time and spectral domain analysis. *IEEE Transactions on Biomedical Engineering*, 2018.

[113] Daniel William Stashuk. Detecting single fiber contributions to motor unit action potentials. *Muscle & nerve*, 22(2):218–229, 1999.

[114] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990.

[115] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2006.

[116] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[117] Claude E Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[118] Jack V Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231, 1996.