

# PAIRWISE ATTRIBUTE NOISE DETECTION ALGORITHM FOR DETECTING NOISE IN SURFACE ELECTROMYOGRAPHY RECORDINGS

by

Gillian Phillips

BSc.EE, University of New Brunswick, 2012

A Thesis Submitted in Partial Fulfilment  
of the Requirements for the Degree of

**Master of Science in Engineering**

In the Graduate Academic Unit of Electrical and Computer Engineering

**Supervisors:** Dawn T. MacIsaac, PhD, Dept. of Electrical & Computer Engineering  
Philip A. Parker, PhD, Prof. Emeritus, Dept. of Electrical & Computer Engineering  
Michael W. Fleming, PhD, Faculty of Computer Science, Assistant Dean

**Examining Board:** Erik Scheme, PhD, Dept. of Electrical & Computer Engineering  
Julian Meng, PhD, Dept. of Electrical & Computer Engineering, Chair  
Scott Bateman, PhD, Faculty of Computer Science

This thesis is accepted by the  
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

May, 2016

© Gillian Phillips, 2016

## **Abstract**

The focus of this work was to modify an existing algorithm originally designed for data mining software metrics and evaluate its usefulness as a quality assessment tool for surface electromyography (sEMG) signals. The pairwise attribute noise detection algorithm (PANDA) was configured to distinguish between clean and noisy sEMG signals. Multiple testing was performed to find the most effective configuration for contamination detection. Data contaminated with power line interference, motion artifact, saturation and combinations of the three were studied. Both simulated and recorded data were used in the configuration and testing stages of this work. PANDA was found to be able to detect low levels of contamination (SNRs of 3 – 17 dB, depending on the type of noise) with high sensitivity (100%). After verifying PANDA's effectiveness, the algorithm was compared to a one-class support vector machine (SVM) designed for the same purpose. For all types of noise, PANDA was more sensitive than the SVM.

## **Acknowledgements**

I would first like to express my gratitude to Dr. Dawn MacIsaac, Dr. Philip Parker, and Dr. Michael Fleming without whom this work would not have materialized. Their ongoing support and patience will be forever appreciated. Many thanks also are owed to Yiyang Shi who helped in many ways.

I would next like to thank my partner, Duncan, who has offered me nothing but unwavering support through this process; I am so lucky! Next, I would like to acknowledge my parents who also have been extremely supportive, both emotionally and through the provision of endless loads of laundry and meals.

I would be remiss not to mention all of the participants who volunteered to flex their muscles for me – thanks for the gun shows, everyone!

Also, thank you to Shelley, Denise, and Karen for putting up with me for so long.

And last, but not least, many thanks to my colleagues and superiors at NB Power for supporting me through my studies.

## Table of Contents

|                                                                            |     |
|----------------------------------------------------------------------------|-----|
| Abstract .....                                                             | ii  |
| Acknowledgements .....                                                     | iii |
| Table of Contents .....                                                    | iv  |
| List of Figures.....                                                       | vi  |
| List of Tables.....                                                        | ix  |
| Chapter 1.            Introduction.....                                    | 1   |
| Chapter 2.            Quality Assessment in Surface Electromyography ..... | 4   |
| 2.1 Introduction .....                                                     | 4   |
| 2.1-1    Surface Electromyography.....                                     | 4   |
| 2.1-2    sEMG Applications.....                                            | 5   |
| 2.1-3    sEMG Contamination .....                                          | 6   |
| 2.2 Quality Assessment of sEMG .....                                       | 9   |
| 2.2-1    CleanEMG .....                                                    | 9   |
| 2.2-1a. Single-Source Solutions.....                                       | 10  |
| 2.2-1b. Any-Source Solutions.....                                          | 12  |
| 2.3 PANDA - Pairwise Attribute Noise Detection Algorithm .....             | 13  |
| 2.3-1    PANDA as a Ranking Algorithm .....                                | 13  |
| 2.3-2    PANDA as a Classifier .....                                       | 17  |
| Chapter 3.            Developing PANDA for use in sEMG .....               | 18  |
| 3.1 Factors under investigation .....                                      | 18  |
| 3.1-1    Potential EMG Features.....                                       | 18  |
| 3.1-2    Baseline data .....                                               | 21  |
| 3.1-1    Boundary decision criteria .....                                  | 22  |
| 3.1-2    Algorithm settings.....                                           | 23  |
| 3.2 Noise types under investigation .....                                  | 25  |
| 3.3 Configuration testing.....                                             | 27  |
| 3.4 Results of investigation .....                                         | 28  |
| 3.5 Proposed configuration .....                                           | 31  |
| Chapter 4.            PANDA in use with simulated baseline data.....       | 32  |
| 4.1 Methods.....                                                           | 32  |
| 4.1-1    Recorded data collection.....                                     | 32  |
| 4.1-2    Record contamination .....                                        | 34  |
| 4.1-3    Testing method.....                                               | 35  |
| 4.2 Results .....                                                          | 35  |
| 4.3 Discussion .....                                                       | 38  |
| Chapter 5.            PANDA in use with recorded baseline data .....       | 41  |

|                                      |    |
|--------------------------------------|----|
| 5.1 Purpose .....                    | 41 |
| 5.2 Methods .....                    | 41 |
| 5.2-1 Recorded data collection ..... | 41 |
| 5.2-2 Record contamination .....     | 42 |
| 5.2-3 Testing method .....           | 43 |
| 5.3 Results .....                    | 49 |
| 5.4 Discussion .....                 | 55 |
| Chapter 6. Comparison to SVM .....   | 57 |
| 6.1 Purpose .....                    | 57 |
| 6.2 Method .....                     | 58 |
| 6.2-1 Recorded data collection ..... | 58 |
| 6.2-2 Record contamination .....     | 58 |
| 6.2-3 Testing method .....           | 58 |
| 6.3 Results .....                    | 59 |
| 6.4 Discussion .....                 | 64 |
| Chapter 7. Conclusion .....          | 67 |
| 7.1 Recommendations .....            | 68 |
| References .....                     | 71 |
| Appendix .....                       | 74 |
| Curriculum Vitae .....               |    |

## List of Figures

|                                                                                                                                                                                                                                                                                                                |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1: Typical clean sEMG and contaminated sEMG (power line interference SNR = 4.77dB) .....                                                                                                                                                                                                                | 6  |
| Figure 2: Typical sEMG Acquisition Setup.....                                                                                                                                                                                                                                                                  | 8  |
| Figure 3: Clustering of feature vectors. Expected values for the $n^{\text{th}}$ signal are collected in $f_k(n)$ , a vector of mean feature values taken across the cluster of which $f(n)$ is a member in the $k^{\text{th}}$ cluster set. Standard deviation vectors, $f_k(n)$ may also be calculated. .... | 15 |
| Figure 4: Binning Methods .....                                                                                                                                                                                                                                                                                | 24 |
| Figure 5: Process used to add contaminants a) Power line interference SNR 3dB b) Motion artifact SNR 3dB c) Saturation 9.2dB .....                                                                                                                                                                             | 26 |
| Figure 6: Process to add simulated noise to clean test signals.....                                                                                                                                                                                                                                            | 27 |
| Figure 7: Noise Factors of baseline signals and signals PL0 to PL7 (feature set combination 3, $N = 600$ , $L = 5$ ). ....                                                                                                                                                                                     | 29 |
| Figure 8: Testing results of initial investigation of PANDA with simulated baseline and test data. Noise type is power line interference. ....                                                                                                                                                                 | 30 |
| Figure 9: Data collection tool .....                                                                                                                                                                                                                                                                           | 33 |
| Figure 10: Process used to add contaminants a) Power line interference SNR 3dB b) Motion artifact SNR 3dB c) Saturation 3dB .....                                                                                                                                                                              | 35 |
| Figure 11: Simulated baseline data testing with recorded test data. Simulated $F_N$ in blue, clean recorded $F_N$ in green, and contaminated (PL7) recorded $F_N$ in red.....                                                                                                                                  | 36 |
| Figure 12: Simulated baseline data with recorded test data Simulated $F_N$ in blue, recorded $F_N$ in green, and contaminated (MA7) recorded $F_N$ in red. ....                                                                                                                                                | 37 |
| Figure 13: Simulated baseline data with recorded test data. Simulated $F_N$ in blue, recorded $F_N$ in green, and contaminated (SAT6.7) recorded $F_N$ in red. ....                                                                                                                                            | 37 |
| Figure 14: Process used to add contaminants a) PL3 b) MA3 c) SAT3 d) PL-MA4.3, which is equivalent to the addition of PL6 and MA6 e) PL-MA-SAT4.3, which is equivalent to the addition of PL6, MA6, and SAT18.8.....                                                                                           | 43 |
| Figure 15: Probability curve .....                                                                                                                                                                                                                                                                             | 46 |
| Figure 16: ROC testing criteria.....                                                                                                                                                                                                                                                                           | 46 |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                      |    |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 17: ROC investigation of a) bin number (for a SD multiplier of 3) and b) standard deviation boundary multiplier (for a bin number of 5) for the addition of power line interference to sEMG at varying SNRs .....                                                                                                                                                                                                                             | 48 |
| Figure 18: Sensitivity of clean data detection in PANDA .....                                                                                                                                                                                                                                                                                                                                                                                        | 50 |
| Figure 19: a) Classification of sEMG contaminated with power line interference (Recall $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with PL7 (red). c) Power spectrums of clean and contaminated records. ....                                                                                                                                                                                        | 51 |
| Figure 20: a) Classification of sEMG contaminated with motion artifact (Recall $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with MA7 (red). c) Power spectrums of clean and contaminated records. ....                                                                                                                                                                                                | 52 |
| Figure 21: a) Classification of sEMG contaminated with saturation (Recall $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with B (red). c) Power spectrums of clean and contaminated records. ....                                                                                                                                                                                                       | 53 |
| Figure 22: a) Classification of sEMG contaminated with power line interference and motion artifact (recall $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with PL3 and MA3 (red), equivalent to PL-MA0.9. c) Power spectrums of clean and contaminated records. ....                                                                                                                                    | 54 |
| Figure 23: a) Classification of sEMG contaminated with power line interference, motion artifact, and saturation (recall $P(FA)=9.6\%$ ). The blue and black represent additions of 1% and 0.5% of saturation, respectively. b) Time domain of clean sEMG record (black) and sEMG record contaminated with PL20, MA20, and SAT18.5 (or 1% saturation) (red), equivalent to PL-MA-SAT16.16. c) Power spectrums of clean and contaminated records. .... | 55 |
| Figure 24: Results of testing the baseline data set (blue points) and clean data set (green points) with PANDA .....                                                                                                                                                                                                                                                                                                                                 | 60 |
| Figure 25: Results of testing clean test data set in the SVM .....                                                                                                                                                                                                                                                                                                                                                                                   | 60 |
| Figure 26: Comparison of any-source solutions testing power line interference (recall $P(FA) = 3.0\%$ ). ....                                                                                                                                                                                                                                                                                                                                        | 61 |
| Figure 27: Comparison of any-source solutions testing motion artifact (recall $P(FA) = 3.0\%$ ).....                                                                                                                                                                                                                                                                                                                                                 | 62 |

Figure 28: Comparison of any-source solution testing saturation (recall  $P(\text{FA}) = 3.0\%$ ).  
.....62

Figure 29: Comparison of any-source solutions testing the combination of power line  
interference and motion artifact (recall  $P(\text{FA}) = 3.0\%$ ).....63

Figure 30: Comparison of any-source solution testing the combination of power line  
interference, motion artifact, and saturation (recall  $P(\text{FA}) = 3.0\%$ ).....63

**List of Tables**

Table 1: Causes and types of sEMG contamination .....8

Table 2: Conventional features [15] .....19

Table 3: Non-traditional features .....20

Table 4: MyoSim parameters.....22

Table 5: Noise Source types and descriptions .....25

## **Chapter 1. Introduction**

The focus of this work was to modify and evaluate the end use of a noise detection algorithm, originally designed for data mining software metrics, to detect contamination in surface electromyography (sEMG) signals. The algorithm is known as the pairwise attribute noise detection algorithm (PANDA) [1]. This work is part of a larger project – CleanEMG - focused on amalgamating algorithms and techniques to identify and quantify contamination found in sEMG signals [2]. SEMG is widely used in a variety of applications including athletic training, diagnoses of neuromuscular diseases, control of prostheses, etc. [3]. It is imperative that the quality of the signals be high to provide accurate results and control in most applications. Many types of contamination exist that can degrade the quality of sEMG signals, including, but not limited to, power line interference, motion artifact, and instrumentation saturation.

There are methods that exist to detect contamination in sEMG, but most are limited to the detection of one type of noise. The approach proposed in this work has the added potential for detecting contamination in sEMG regardless of the type of noise or whether one or multiple types of noise are contributing to the contamination. To use PANDA for sEMG quality assessment, it is first tuned using a specified set of features estimated from a clean baseline data set (similar to training data). Then, the algorithm accepts a test record, and classifies it as either clean or contaminated. Classification is based on a noise factor calculated for each example presented to PANDA. The baseline examples are used to establish a noise factor range for clean examples. Test records with noise factors which fall outside that range are classified as noisy.

An overview of quality assessment in sEMG is explored in Chapter 2. Subsequent chapters detail work conducted to meet the following objectives:

1. To develop PANDA for use in sEMG quality analysis: To meet this objective, suitable algorithmic parameters for PANDA were established along with a suitable feature set. This was accomplished by using simulated data as both the baseline data set and test records and is detailed in Chapter 3. These preliminary results, based on simulated data, indicated that PANDA was a viable alternative for quality assessment in sEMG.

2. To test the PANDA configuration using recorded data as the test data set: It is difficult to guarantee the cleanliness of data collected from a lab environment. For this reason, the work aimed to test whether a set of simulated baseline signals could be used to tune PANDA for detection of contamination in recorded data. This is detailed in Chapter 4. The results of this exploration indicated that the current state of the simulation tool produces data which is not viable for use with PANDA. This result instigated objective 3, which follows.

3. To test PANDA with recorded data employed as both the baseline and test data sets: During this portion of the work, the configuration of PANDA was refined to replace simulated data with recorded data as the baseline data; this is detailed in Chapter 5. The results of this exploration significantly improved performance and re-validated PANDA as a quality assessment tool for sEMG.

4. To compare the sensitivity of PANDA with another similar noise detection algorithm: To meet this objective, PANDA was compared to a support vector

machine designed for the same purpose as detailed in Chapter 6. The results of this exploration indicated that PANDA outperforms the SVM as a quality assessment tool for sEMG.

This work investigated three main types of noise – power line interference, motion artifact, and instrumentation saturation, both individually and collectively.

## **Chapter 2. Quality Assessment in Surface Electromyography**

### **2.1 Introduction**

#### **2.1-1 Surface Electromyography**

Signals sent from the nervous system elicit small electrical signals from skeletal muscles, which result in muscle contractions. The electrical signals that trigger contractions are known as myoelectric signals and measurement of these signals is accomplished through a process known as electromyography (EMG). As such, the data recorded during EMG is often referred to as an EMG signal. Two EMG techniques exist, one that is invasive and another that is non-invasive. The latter is referred to as surface electromyography (sEMG) and is the focus of this work.

Measuring electrodes are used when collecting EMG data. The measuring electrodes are placed on the surface of the skin, over the belly of the muscle of interest during sEMG. Alternatively, invasive or intramuscular EMG uses a needle electrode inserted into the muscle belly. The invasive technique yields signals with higher fidelity than its non-invasive counterpart, but it is more technically demanding and requires considerable training to do properly. The process is also uncomfortable for participants and increases the risk of harm due to the potential for infection. The decreased signal fidelity from sEMG is inherent to the separation between the measuring electrode and the muscle of interest. Skin and tissue fill this gap and behave like a low-pass filter. Also, surface electrodes are likely to pick up interference from other muscles' signals in the vicinity and are more sensitive to muscle activity closest to the surface. Nevertheless, sEMG yields signals suitable for use in many applications [3] and was chosen as the main interest for this work due to its ease of use.

## **2.1-2 sEMG Applications**

Ergonomics, rehabilitation, monitoring and diagnosis of neuromuscular diseases, athletic training, and assistive device control are all areas in which sEMG has made contributions [3].

For applications related to rehabilitation, collection and analysis of data is often performed by a therapist to help clients better conform to their rehabilitation protocols. For instance, Bolek [4] reported on a study where sEMG was used by physiotherapists to positively reinforce proper movements of 16 pediatrics patients receiving treatment for various movement disorders, such as cerebral palsy.

In the field of ergonomics, researchers and consultants hold much of the responsibility of sEMG collection and analysis. For instance, graduate students at the University of New Brunswick Occupational Biomechanics Lab used sEMG to investigate the physical demands incurred by local police officers while in their cruisers [5]. The police officers were fitted with sEMG collection devices and the data was later examined by the researchers. Some tools, however, allow for individuals to collect and examine their own sEMG data. One such device is the Pocket Ergometer [6], which helps users to self-identify stressed muscles using surface electrodes placed over various muscles.

An emerging application in the device control field may generate more ubiquitous use of sEMG. Engineering students at the University of Waterloo recently developed a gesture control armband called MYO<sup>TM</sup>, which uses sEMG to operate. The armband can be synced, for example, to an iPad and the user can operate it with a flick of the wrist or a clench of the fist [7].

SEMG has many useful applications and, as evidenced through the examples, the collection and/or analysis of sEMG data can be completed by a spectrum of individuals – from trained professionals to end-users with little to no knowledge of EMG or the collection/analysis process. Even clinicians trained in collection rarely have the expertise required to reliably assess the data they collect in terms of its quality, which is not a trivial task. A typical sEMG recording resembles random noise, as depicted in Figure 1a). Figure 1b) shows the clean signal from Figure 1a) with added medium intensity power line interference contamination,  $\text{SNR} = 4.77\text{dB}$ . As exemplified by these figures, it is difficult, through visual inspection, to distinguish between clean and noisy signals. Automated quality assessment, therefore, is a definite asset in sEMG-based systems.

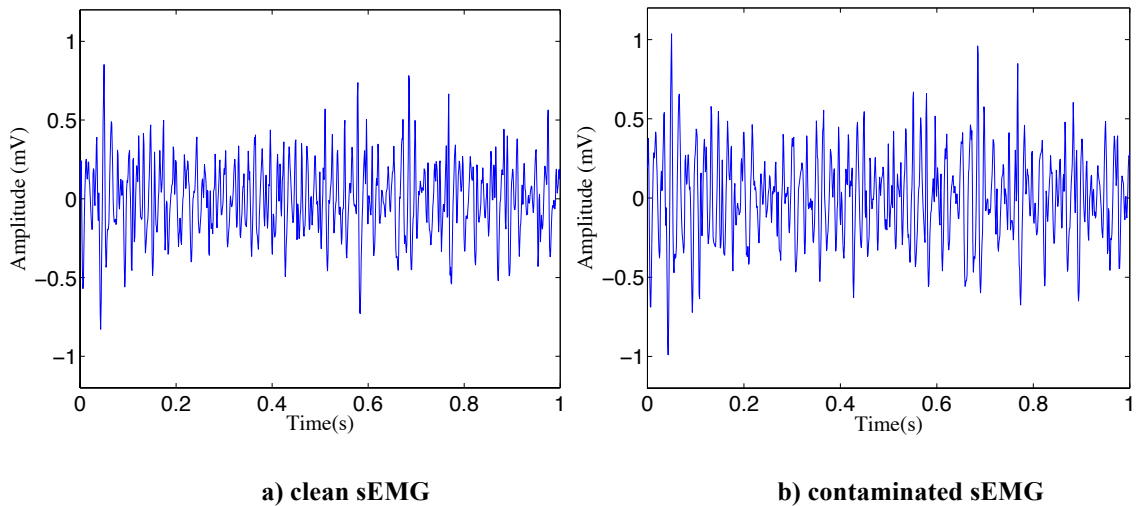


**Figure 1: Typical clean sEMG and contaminated sEMG (power line interference  $\text{SNR} = 4.77\text{dB}$ )**

### 2.1-3 sEMG Contamination

A signal's usefulness is often correlated to its quality, which can be hard to guarantee. In [8], Grönlund et al. discuss the matter of low-quality sEMG signals and a number of

offending sources of noise. Grönlund refers to the work conducted by Clancy et al. [9] and Huigen et al. [10], among others, who indicated that acquisition of sEMG data is multi-faceted and must be completed carefully; otherwise, recordings are at risk of contamination from various noise sources, which can reduce their quality. Acquisition considerations include:

- The skin where the electrode will be placed must be properly prepared. This entails sloughing off the dead skin cells, cleaning the area with an alcohol wipe and rubbing with electrode gel.
- The electrode must be placed in the proper location for the muscle in question. There are guidelines, such as those developed by SENIAM, that can be followed [11].
- The electrodes must also be adhered to the skin properly to avoid lift or movement during data collection. An adhesive sticker, medical tape or electrode arm-band can be used.
- The instrumentation should be chosen and tuned appropriately.

Figure 2 illustrates a typical sEMG acquisition setup. Wired or wireless electrodes, like the Delsys TRIGNO™ electrodes, can be used. Some electrodes, like those of the Delsys System, have the instrumentation amplifiers built in. Systems with all the instrumentation built-in may be easier to use since they are pre-configured, but systems that allow users to set instrumentation parameters like gains and cut-off frequencies may be more robust.

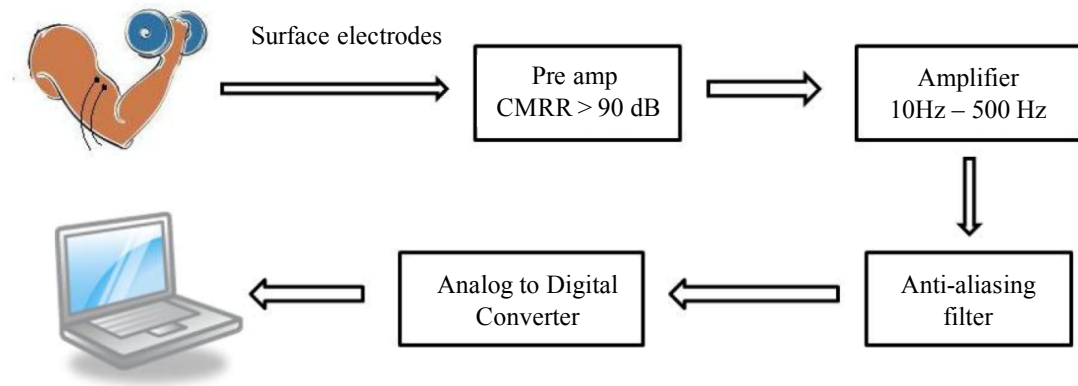


**Figure 2: Typical sEMG Acquisition Setup**

Following an appropriate acquisition procedure is necessary, but a noise-free (clean) recording is still difficult to acquire. Many sources of noise can contaminate sEMG data, including those listed below:

**Table 1: Causes and types of sEMG contamination [2][9][10][12][13] [14]**

| Cause of contamination                                                  | Type of contamination                                                   |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------|
| <b>Amplifier saturation</b>                                             | EMG clipping                                                            |
| <b>Analog-to-digital converter (ADC) over-ranging</b>                   | EMG clipping<br>Quantization noise                                      |
| <b>Insufficient pre-amplification (low common mode rejection ratio)</b> | Power line interference<br>EMG Clipping <sup>†</sup>                    |
| <b>Poor electrode contact and cable movement</b>                        | Motion artifact<br>Power line interference<br>EMG Clipping <sup>†</sup> |
| <b>Physiological interference</b>                                       | ECG interference<br>Muscle cross-talk                                   |

<sup>†</sup> Insufficient pre-amplification or poor electrode contact can cause amplifier saturation, leading to EMG clipping, due to reduced common mode signal rejection

Some noise is expected in all sEMG recordings, and its presence does not necessarily preclude the utility of sEMG. While the threshold for acceptable contamination may be application specific, algorithms can be used to determine when a recording is contaminated beyond the threshold for use. Such algorithms are especially useful if they

can be integrated into data collection, in order to immediately throw away any bad data and collect new signals. They can also be used in offline processing, when redundant data is available, to choose only the best recordings for further processing.

## **2.2 Quality Assessment of sEMG**

### **2.2-1 CleanEMG**

The CleanEMG research project [2] is an ongoing initiative established to provide open source quality assessment solutions for sEMG. The project focuses on contamination resulting from inadequate instrumentation and measurement set-ups, and interference. The project emphasizes the need for integration of an automatic signal quality assessment tool into sEMG acquisition systems. This is especially crucial for clinicians or end-users who collect/analyze/use sEMG but may not possess the skills of an instrumentation or biosignals expert or have the time to judge the quality of sEMG data.

Many techniques are emerging to detect, identify, and/or quantify contamination in sEMG records. The most common approach characterizes the type of noise [12] [13] [15][16]. If a measurable characterization of the noise type in question can be achieved, the noise can then be detected in the record and identified, and in some cases quantified or removed. Alternatively, a relatively untapped approach characterizes clean sEMG signals to differentiate them from noisy signals. A criterion for clean signals is specified and any signal not meeting the criterion is considered contaminated. This approach has the advantage of not requiring knowledge of the type of contamination in the sEMG signal. It also may work better in cases for which multiple noise sources are present, possibly influencing each other's characteristics.

### 2.2-1a. Single-Source Solutions

To date, most of the CleanEMG methods being developed focus on a particular type of noise, attempting to characterize it sufficiently to detect it within an sEMG recording.

Abser et al. [12] sought to quantify power line interference using a method developed in previous work by Mewett et al. [16]. Mewett et al. used spectral interpolation to reduce power line interference in sEMG data, which is known to exist at 50Hz or 60Hz. The method transforms the sEMG signal into the frequency domain and interpolates the amplitude spectrum at the noted frequencies. Interpolation is used instead of notch-filtering to maintain the integrity of the original sEMG signal, while mitigating the power line interference.

Abser et al. [12] retained both the cleaned signal and the estimated noise in their work so that they could quantify the power line interference by calculating the Signal to Noise Ratio (SNR) according to:

$$SNR = \frac{Power(cleaned\ signal)}{Power(estimated\ noise)}$$

Using simulations to control level of contamination, the work successfully demonstrated power line interference quantification in sEMG.

In [14], Fraser et al. investigated a new method to remove electrocardiogram (ECG) artifacts from sEMG. The proposed method used moving averages to estimate the ECG contaminants. Two moving averages were used – one to estimate the higher frequency components of the contamination and the other to estimate the lower frequency

components. The two averages were combined to estimate the ECG form, which was then subtracted from the contaminated signal.

This method was compared to the template subtraction method [17], which also shows promising performance when compared to other ECG subtraction techniques. Template subtraction techniques require a reference signal of the ECG contaminant. In [17], this was accomplished by recording the muscle under examination when it was relaxed to attain an sEMG-free recording of the contaminating ECG signal. A thresholding technique was then used to fully distinguish the ECG signal from the sEMG-free recording. The ECG signal could then be subtracted from the contaminated sEMG signal to produce a clean recording.

The comparison in [7] demonstrated that the moving average method performed better than template subtraction methods for contamination with lower SNR values ( $< 0$  dB with a testing range from -8dB to 8dB). Fraser et al. also pointed out that their method works better for real-time applications because it does not require the relaxed sEMG recording. However, the template subtraction method did outperform the moving average method for contaminants with higher SNR values ( $> 0$  dB).

In another related work, Fraser et al. [13] investigated methods to detect and quantify three other contaminants of sEMG data: analog-to-digital converter (ADC) clipping, quantization noise, and amplifier saturation. ADC clipping was detected by searching for two consecutive maxima or minima in a signal. Given the random nature of sEMG signals, there is low likelihood that two consecutive maxima or minima will occur without the presence of ADC clipping. Quantization noise was detected using a signal-

to-quantization-noise ratio (SQNR) and an estimation of the smallest step size in the sEMG signal. Finally, amplifier saturation was detected through a test of normality. The amplitudes of an sEMG signal should follow a normal distribution, but as saturation in a signal increases, the normal distribution is distorted. The method employed the Pearson correlation coefficient between the sEMG amplitude histogram and the normal probability density function to test for saturation. Through testing with simulated and recorded data (for the saturation test only), all three methods were verified.

### **2.2-1b. Any-Source Solutions**

Recently, CleanEMG researchers have begun to explore any-source solutions, which focus on characterizing the signal, not the noise. Fraser et al. [18] trained a one-class support vector machine (SVM) with data known to be noise-free to determine how well the machine could differentiate between clean and noisy sEMG recordings. Both simulated and recorded data were used in this work and six types of contaminants were investigated, including power line interference, motion artifact, ECG interference, analog-to-digital converter (ADC) clipping, quantization noise, and amplifier saturation.

The team found that the SVM could detect varying levels of contamination depending on the contaminant type. The machine was more sensitive to higher levels of contamination and sensitivity fell as the contamination level decreased; a transition point, where the sensitivity began to decline, was noted for each contaminant type. For power line interference, motion artifact, and ECG interference, the SNR transition points were all found to be less than or equal to 5dB. For ADC clipping, quantization noise, and amplifier saturation, the transition points were all found to be greater than or equal to 12dB.

In an attempt to improve on the current state of any-source solutions, the work reported in this thesis explored a preprocessing data mining algorithm called the Pairwise Attribute Noise Detection Algorithm (PANDA) to differentiate between clean and noisy records. PANDA provides a rank of the input signals in terms of their likelihood to be contaminated. Those ranked highest are considered to be ‘suspect’ recordings. The algorithm is borrowed from previous work completed by software engineers [1][19] to discern suspect data in software development process monitoring. In this work, we modified and implemented PANDA to essentially data mine sEMG signals.

## **2.2-2 Beyond detection and quantification**

CleanEMG [2] focuses on the detection and quantification of contamination in sEMG recordings, primarily for use during data collection, to guide proper set up and discard low quality recordings. Because of the well-documented evidence of noise in sEMG [8] [10] plenty of work has also been done surrounding the removal of the contaminants after collection. Examples of this include: in [20], the sEMG recording is filtered to remove movement artifact and baseline noise contamination; in [16], power line interference is removed from the sEMG recording by a spectrum interpolation method; in [17], researchers employ a subtraction method to remove ECG components from the sEMG recording; in [21], the sEMG recordings are de-noised using wavelet analysis; and in [22], multiple noise sources are removed through a neural network analysis.

## **2.3 PANDA - Pairwise Attribute Noise Detection Algorithm**

### **2.3-1 PANDA as a Ranking Algorithm**

The purpose of this work was to investigate the performance of PANDA in detecting noisy sEMG recordings. As its name implies, PANDA relies on a pairwise comparison

of characteristic attributes defined for a data set. In the context of sEMG quality analysis, the data set is made up of a set of baseline sEMG signals known to be noise-free, together with a signal or signals under investigation. A worked example of PANDA is included in the Appendix to aid the reader in discerning what is happening throughout the following description.

At the start of PANDA, a set of  $M$  features (attributes) is calculated for each signal (observation) in a data set of  $N$  signals to produce  $N$  feature vectors:

$$f^{(n)} = \begin{bmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_M^{(n)} \end{bmatrix} \text{ for } n = 1, 2, 3, \dots N \quad (1)$$

The algorithm works by trying to identify signals with one or more feature value that deviates from what is expected based on other similar signals in the data set. If the data set were strictly homogenous (all the signals are known to be recorded from the same location and from the same person under identical conditions – constant force, constant joint angle, stationary firing statistics, no fatigue, etc), then all of the signals in the data set could be considered similar, and the expected value of any feature could be estimated by the mean of that feature across the data set. When this is not the case, some criteria can be established to cluster the signals into groups of similar signals. PANDA uses partitioned bins of feature values to cluster similar signals, as depicted in Figure 3.

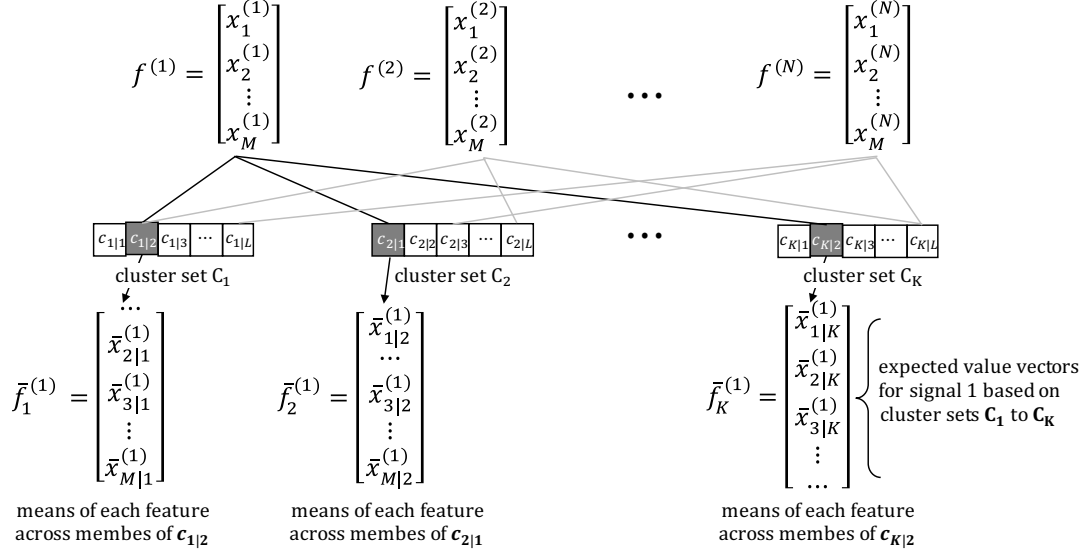


**Figure 3: Clustering of feature vectors.** Expected values for the  $n^{\text{th}}$  signal are collected in  $\bar{f}_k^{(n)}$ , a vector of mean feature values taken across the cluster of which  $f^{(n)}$  is a member in the  $k^{\text{th}}$  cluster set. Standard deviation vectors,  $\bar{\sigma}_k^{(n)}$  may also be calculated.

No a-priori information is assumed about the features. Instead, each feature is simply partitioned into a set of  $L$  contiguous bins and signal clusters are established based on membership in the bins. Each of the features has its own set of bins, yielding  $K=M$  sets of  $L$  signal clusters,  $C_k$ :

$$C_k = \begin{bmatrix} c_{k|1} \\ c_{k|2} \\ \vdots \\ c_{k|L} \end{bmatrix} \text{ for } k = 1, 2, 3, \dots K(=M) \quad (2)$$

Each cluster set contains all  $N$  feature vectors, clustered according to the values of their  $m^{\text{th}}$  feature. For a given cluster set  $C_k$ , the cluster of which the feature vector of interest is a member,  $c_{k|l} \ni f^{(n)}$ , is considered to be the group of similar vectors for that vector, according to the feature that was used to establish the clusters (i.e. the clustering feature).

Taking the mean for each of the other features across the group of similar vectors yields an expected value of those features,  $\bar{x}_{m|k}^{(n)}$ , according to the clustering feature. That is, an expected value for the  $m^{\text{th}}$  feature of the  $n^{\text{th}}$  vector can be estimated by taking the mean of that feature across the vectors clustered together with the  $n^{\text{th}}$  vector according to the  $k^{\text{th}}$  feature. Since this mean is calculated for all clusters sets with  $k \neq m$ ,  $K-1$  expected values for each signal (and each feature) are generated. Standard deviations  $\sigma_{m|k}^{(n)}$  for the  $K-1$  groups can also be calculated. Thus, in Figure 3,  $\bar{f}_k^{(n)}$  represents the vector of feature means for the  $n^{\text{th}}$  signal according to the  $k^{\text{th}}$  feature. Similarly, a vector of feature standard deviations is delineated by  $\bar{\bar{f}}_k^{(n)}$ .

PANDA uses the expected values to quantify the deviation from what is expected of each feature for a given signal by taking the difference between the feature values and their corresponding expected values. To normalize the deviations, they are divided by the standard deviations. Since there are  $K-1$  expected values for each feature (and signal), this process yields a deviation matrix  $\Delta^{(n)}$  for each signal as delineated in (3), where  $\bar{f}^{(n)}$  represents a matrix collection of  $\bar{f}_1^{(n)}$  to  $\bar{f}_K^{(n)}$  and  $\bar{\bar{f}}^{(n)}$  represents the matrix collection of standard deviation vectors:

$$\Delta^{(n)} = \frac{|\bar{f}^{(n)} - f^{(n)}|}{\bar{\bar{f}}^{(n)}} = \frac{|\begin{bmatrix} \bar{f}_1^{(n)} & \bar{f}_2^{(n)} & \dots & \bar{f}_K^{(n)} \end{bmatrix} - f^{(n)}|}{\begin{bmatrix} \bar{\bar{f}}_1^{(n)} & \bar{\bar{f}}_2^{(n)} & \dots & \bar{\bar{f}}_K^{(n)} \end{bmatrix}} \quad (3a)$$

$$= \left\| \begin{bmatrix} \bar{x}_{1|1}^{(n)} & \bar{x}_{1|2}^{(n)} & \bar{x}_{1|3}^{(n)} & \dots & \bar{x}_{1|K}^{(n)} \\ \bar{x}_{2|1}^{(n)} & \bar{x}_{2|2}^{(n)} & \bar{x}_{2|3}^{(n)} & \dots & \bar{x}_{2|K}^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{x}_{M|1}^{(n)} & \bar{x}_{M|2}^{(n)} & \bar{x}_{M|3}^{(n)} & \dots & \bar{x}_{M|K}^{(n)} \end{bmatrix} - \begin{bmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_M^{(n)} \end{bmatrix} \right\| \left/ \begin{bmatrix} \bar{\bar{x}}_{1|1}^{(n)} & \bar{\bar{x}}_{1|2}^{(n)} & \bar{\bar{x}}_{1|3}^{(n)} & \dots & \bar{\bar{x}}_{1|K}^{(n)} \\ \bar{\bar{x}}_{2|1}^{(n)} & \bar{\bar{x}}_{2|2}^{(n)} & \bar{\bar{x}}_{2|3}^{(n)} & \dots & \bar{\bar{x}}_{2|K}^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{\bar{x}}_{M|1}^{(n)} & \bar{\bar{x}}_{M|2}^{(n)} & \bar{\bar{x}}_{M|3}^{(n)} & \dots & \bar{\bar{x}}_{M|K}^{(n)} \end{bmatrix} \right. \quad (3b)^\dagger$$

---

<sup>†</sup>In 3b, division represents an element-by-element operation, not matrix division

The total deviation per feature for a given signal,  $E_m^{(n)}$  can be calculated simply by summing across the rows of the deviation matrix.:

$$E_m^{(n)} = \sum_{k=1}^K \Delta_{mk} \quad (4)$$

Finally, the sum of deviations across features is used as the noise factor,  $F_N$ , for each signal:

$$F_N = \sum_{m=1}^M E_m^{(n)} \quad (5)$$

To date, researchers ranked the  $F_N$  values and then left it up to an analyst to decide which (if any) of the top-ranked observations would be considered too noisy for their application.

### **2.3-2 PANDA as a Classifier**

For this work, PANDA was used as a classifier. This was accomplished by first calculating the  $F_N$  values for the baseline data set and then establishing a boundary based on these values. The  $F_N$  values of the test data set were then found and, based on whether they fell inside or outside of the boundary, they were determined to be clean or contaminated; clean data fell within the boundary and noisy data fell outside of the boundary.

## **Chapter 3. Developing PANDA for use in sEMG**

To explore how to use PANDA for use in sEMG applications, a set of clean sEMG signals was simulated, and attributes that characterized the sEMG were identified. Then, noisy sEMG signals were simulated and PANDA was employed to generate a Noise Factor ( $F_N$ ) for each of the clean and noisy signals. Performance of the algorithm's ability to discern between clean and noisy sEMG data was evaluated based on its ability as a classifier to quantify a boundary between clean and noisy signals with precision. Performance was compared across a variety of configurations involving the factors as described below.

### **3.1 Factors under investigation**

#### **3.1-1 Potential EMG Features**

The fundamental input to PANDA is a set of attributes used to characterize the data. Eight features were selected for investigation in this work to characterize sEMG. Each feature is defined in terms of a sEMG signal,  $x(n\Delta t)$ , as described below, where  $\Delta t = T/N$  is the sampling interval ( $T$  = length of the signal in seconds,  $N$  = number of samples in signal). The first four (delineated in Table 2) are conventional features that have already been shown to perform well in other sEMG-based applications [3]. The others, while less common, were chosen for their potential to recognize uncharacteristic sEMG and are delineated in Table 3.

**Table 2: Conventional features [18]**

| <i>Feature</i>                   | <i>Description</i>                                                                                                                                                                                                                                             | <i>Formulae</i>                                                                                                                                                                                                                                                           |
|----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Mean Absolute Value (MAV)</i> | The mean of a typical EMG signal is approximately zero, therefore the mean absolute value is found to provide a reasonable value for comparisons. MAV is calculated by averaging the absolute values of each sample.                                           | $MAV = \frac{1}{N} \sum_{n=1}^N  x(n\Delta t) $                                                                                                                                                                                                                           |
| <i>Zero Crossings (ZC)</i>       | The zero crossings attribute indicates the number of sign changes that occur within the signal. The ZC count is augmented each time the signal crosses the zero line (x-axis). A threshold (Th) <sup>†</sup> is set to omit any crossings that are negligible. | $ZC = \sum_{n=1}^{N-1} f((x(n\Delta t) \cdot x((n+1)\Delta t)))$ $+ \sum_{n=1}^{N-1} g\left(\frac{x((n+1)\Delta t)}{x(n\Delta t)}\right)$ $f(x) = \begin{cases} 1, & x < Th \\ 0, & x \geq Th \end{cases}$ $g(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq 0 \end{cases}$ |
| <i>Slope Sign Change (SSC)</i>   | The slope sign change attribute denotes the number of slope sign changes incurred by the signal. Like the ZC attribute, a threshold (Th) <sup>†</sup> is assigned to omit negligible slope sign changes.                                                       | $SSC = \sum_{n=2}^{N-1} f((x(n\Delta t) - x((n-1)\Delta t))$ $\cdot (x((n+1)\Delta t) - x(n\Delta t)))$ $f(x) = \begin{cases} 1, & x \leq Th \\ 0, & x > Th \end{cases}$                                                                                                  |
| <i>Wavelength (WL)</i>           | The wavelength of the signal is calculated by computing the shortest distance (the hypotenuse) between each adjacent pair of samples and summing the findings.                                                                                                 | $WL = \sum_{n=2}^N \sqrt{(x(n\Delta t) - x((n-1)\Delta t))^2 + \Delta t^2}$                                                                                                                                                                                               |

<sup>†</sup>For this work, the threshold for SSC and ZC was set to Th=0.005.

The features in Table 3 are less common in the literature, but were deemed useful for exploration in this application. Entropy has recently gained traction in other sEMG applications so it was worth including in this work [23]. The 60Hz and 180Hz interference features were included to aid the algorithm in identification of power line interference, a common noise source for sEMG. In the table,  $X(n\Delta f)$  is the frequency domain representation of  $x(n\Delta t)$ , where  $\Delta f = 1/T$  is the frequency resolution. The final

feature, signal to motion artifact ratio, is included for its utility in detecting low frequency noise, such as motion artifact.

**Table 3: Non-traditional features**

| <b>Feature</b>                              | <b>Description</b>                                                                                                                                                                                                                                                                                                  | <b>Formulae</b>                                                                                                                                                                                                                                                                                                                                                                |
|---------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Entropy (EN)</i>                         | Entropy measures the randomness of the amplitude of the input signal. Here, $p_m$ represents the probability of occurrence of the $m^{\text{th}}$ amplitude value ( $A_m$ ). It is estimated by counting the number of times it occurs in the given signal (ie from a histogram of $x(n\Delta t)$ ).                | $EN = - \sum_{m=1}^M (p_m \cdot \log_2(p_m))$ $p_m \approx \frac{\text{count}(A_m)}{N}$                                                                                                                                                                                                                                                                                        |
| <i>60Hz Interference</i>                    | The 60 Hz interference feature is implemented by summing the spectral components in a narrow range around 60 Hz. A range was used to accommodate potential for 60Hz jitter described by Abser et al [12]. In their work they suggest a range between 58-62Hz which was applied here.                                | $Int(60) = \sum_{f=58}^{62} [X(n\Delta f)]^2$                                                                                                                                                                                                                                                                                                                                  |
| <i>180Hz Interference</i>                   | The 180 Hz interference feature is similar to the 60 Hz interference feature, except the frequency range is 178 to 182 Hz.                                                                                                                                                                                          | $Int(180) = \sum_{f=178}^{182} [X(n\Delta f)]^2$                                                                                                                                                                                                                                                                                                                               |
| <i>Signal to Motion Artifact Ratio (SM)</i> | The SM ratio is used to identify low-frequency noise in the signal [24]. It is found by summing the power densities across the entire signal and dividing by the power densities at frequencies less than 20Hz which extend beyond a line drawn from the origin (zero frequency) to the highest mean power density. | $SM = \frac{\sum_{i=0}^N  X(n\Delta f) ^2}{\sum_{i=0}^K  X_{peak}(k\Delta f) ^2}$ <ul style="list-style-type: none"> <li>• where <math>k</math> represents the frequencies <math>\leq 20\text{Hz}</math></li> <li>• where <math>X_{peak}</math> represents the power densities that extend beyond the line drawn from the origin to <math>\max(X(n\Delta f))</math></li> </ul> |

Preliminary exploration indicated that individual features were often more or less sensitive to a particular noise source, but a multi-feature combination was necessary to capture varied or multiple noise sources. Three combinations of the features were explored to determine their influence on PANDA's performance:

- Combination 1: The four conventional features: MAV, ZC, SSC, WL,
- Combination 2: The other four features EN, 60Hz, 180Hz, SM
- Combination 3: All eight features MAV, ZC, SSC, WL, EN, 60Hz, 180Hz, SM.

### **3.1-2 Baseline data**

In the software applications for which PANDA was originally intended, the algorithm was designed to take one dataset, calculate a noise factor for each observation, and rank all observations from most to least noisy. It was up to an analyst to decide which (if any) of the top-ranked observations would be considered too noisy for their application. For the sEMG application, we modified the way PANDA is used; we take a baseline data set that is known to be sufficiently noise-free and use it to establish the expected range of noise factors for clean data. Then, noise factors for one or more test data can be calculated, and data that falls within the established boundary are presumed clean, while those falling outside the boundary are presumed to be noisy. In order to do this, we must first establish a baseline data set. For the initial exploration, we used simulated data so that we could ensure the data was clean.

SEMG data was simulated with Myosim [25], an open-source EMG simulation tool. A set of 1000 one-second signals was modeled from the biceps brachii by setting parameters that are physiologically appropriate, as outlined in Table 4.

**Table 4: MyoSim parameters**

|                     |                                              |                    |
|---------------------|----------------------------------------------|--------------------|
|                     | Number of fibers per motor unit              | 50 to 100          |
|                     | Initial number of motor units                | 50                 |
| Fiber Parameters    | Location of distal fiber termination         | -220 +/- 5 mm      |
|                     | Location of proximal fiber termination       | 180 +/- 5 mm       |
|                     | Innervation point dispersion                 | +/- 5 mm           |
| Fiber Locations     | Range of vertical depth of motor units       | 10 to 30 +/- 3 mm  |
|                     | Range of horizontal alignment of motor units | -10 to 10 +/- 5 mm |
|                     | Radius of limb                               | 40 mm              |
| Conduction velocity | Fiber conduction velocity                    | 4 +/- 2 m/s        |
| Firing Statistic    | Motor unit firing rate (pulses per second)   | 8 * 0.25 Hz        |

The number of records in the baseline data set was explored to determine its influence. Baseline data set sizes of 300, 600 and 1000 observations were tested to establish a suitable configuration.

### **3.1-3 Boundary decision criteria**

Since the sEMG application of PANDA uses a known set of clean signals to establish an expected range of noise factors for clean data, a statistical variation (defined in terms of standard deviation) from the mean noise factor of these signals is useful in defining a boundary between noise factors from clean vs noisy signals. There is a tradeoff between setting the boundary too high or too low. A high boundary will correctly identify more clean signals, but may falsely identify more noisy signals as being clean. Conversely, a

boundary that is set too low will identify more noisy signals as being noisy, but will also incorrectly identify clean signals as being noisy. The balance point for this tradeoff will likely be application specific. For instance, in circumstances when lots of redundant data is available, throwing away some clean records may not be problematic. Alternatively, for applications which are resilient to noise, including a few noisy records may not be problematic. For this work, the balance point is tipped towards the former, thus throwing away some clean data records.

For the initial exploration, the boundary was set at three standard deviations from the mean. Statistically, 99.7% of a normally distributed data set will fall within three deviations from the mean. In later testing, receiver operating characteristic testing was used to investigate high/low boundary setting tradeoffs (which is described in subsequent chapters).

### **3.1-4 Algorithm settings**

A key component to applying PANDA is establishing how to partition each of the features into contiguous bins. In PANDA's original application, Khoshgoftaar and van Hulse [19] used an equal-frequency binning approach, meaning the bins were set so that the same number of observations fell into each [as depicted in Figure 4a], but suggested other approaches would be investigated in their future work. In this work, two types of binning were considered – binning with equal frequency and binning with equal bin width. The latter means that the widths of the bins are equal, so a varying number of observations can fall within each bin [as depicted in Figure 4b].

Figure 4 depicts each binning method through a simple example. There are  $N = 15$  observations ranging in value from 100-200 and  $L = 5$  bins. The equal-frequency method (a) will place  $N/L = 15/5 = 3$  observations in each bin. The equal bin width method will have a bin width =  $(\max_{\text{obs}} - \min_{\text{obs}})/L = (200-100)/5 = 20$ . The observations will fall into the bin for which they are designated, based on the width.

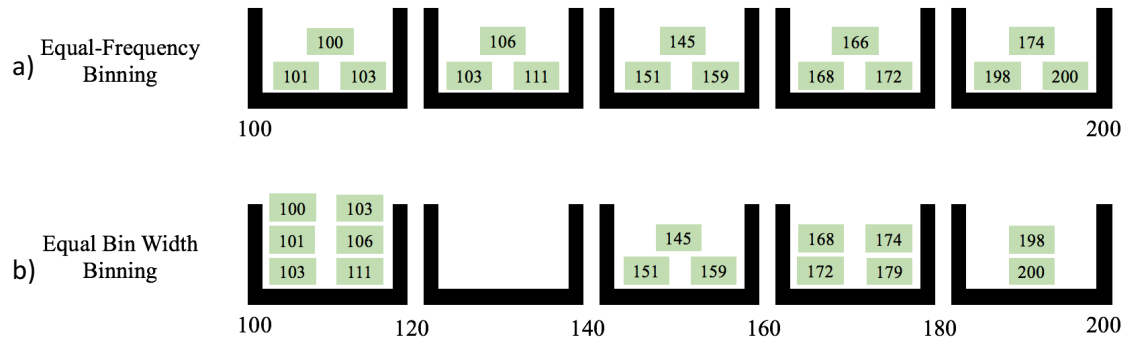


**Figure 4: Binning Methods**

Preliminary ad hoc testing indicated that the equal bin width approach worked well for the sEMG application, so we continued using that approach. While we considered equal-frequency binning, we eventually decided not to use it because we had a viable alternative, and equal-frequency binning was hindered by a significant weakness. It would sometimes result in irregular groupings of attributes. For example, a group of similar feature values could arbitrarily be separated because a bin had met its capacity; this can be observed in Figure 4 where two observations both have a value of 103. In the equal bin width method, the two observations are binned together and would always be, using this method. In the equal frequency method, the two observations are separated into different bins due to the frequency requirement of each bin. The nature of PANDA

is to compare groupings of signals based on the similarity of their feature values; the equal frequency method can work in opposition to this as just described.

Sets of 5, 8, and 15 bins were explored to establish a suitable bin width configuration.

### 3.2 Noise types under investigation

Three noise types were selected for investigation with PANDA as listed in Table 5. Each is regularly featured in the CleanEMG project [2] and procedures already developed as part of that project were followed to simulate the noise types. Simulations were used to control noise level so that its influence on performance could be examined. Simulated noise was added to simulated data for analysis.

**Table 5: Noise Source types and descriptions**

| Noise Source            | Description                                                                                                                                       |
|-------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| Power line interference | The 50-60 Hz disturbance resulting from electromagnetic coupling.                                                                                 |
| Motion artifact         | Low frequency disturbance caused by electrode and/or cable movement during acquisition.                                                           |
| Saturation              | Disturbance throughout the sEMG caused when a nonlinearity is introduced due to the signal output approaching the amplifier power supply voltage. |

Three types of noise were tested: power line interference (PL); motion artifact (MA); and instrumentation saturation (SAT). This work describes a contaminated record with the naming convention of [type of contamination][SNR amount in dB]. For example, a record contaminated with power line interference at an SNR of 7 dB is referred to as PL7. Figure 5 illustrates the process used to add the different contaminants.

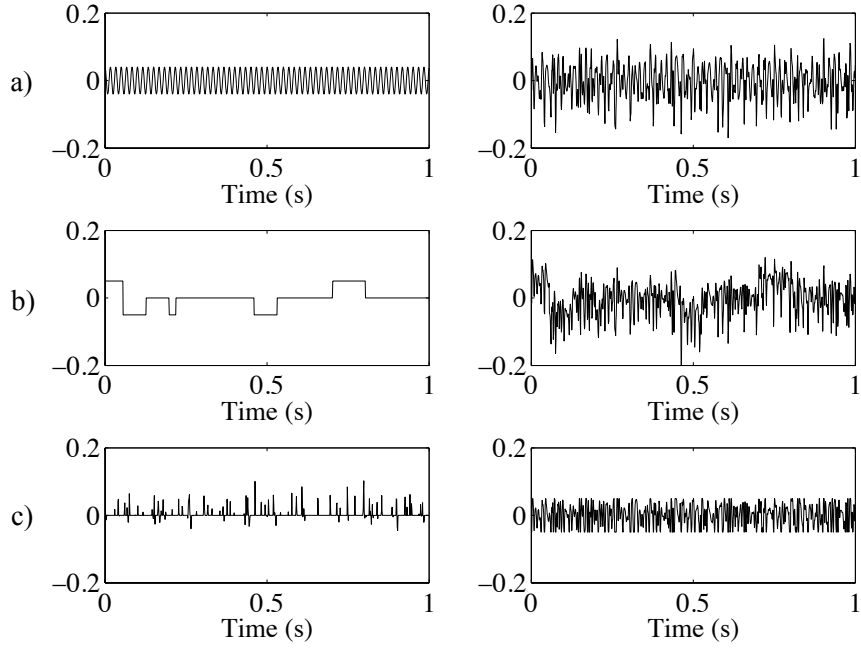


**Figure 5: Process used to add contaminants a) Power line interference SNR 3dB b) Motion artifact SNR 3dB c) Saturation 9.2dB**

To add PL, a sinusoid with random frequency (59.5-60.5Hz,  $\Delta f=0.5\text{Hz}$ ) and phase  $\varphi = 0$  was added to each record. Amplitude of the sinusoid was adjusted to achieve SNR values of 0 - 7dB.

To add MA, a pulse train with randomly scattered negative and positive pulses was added to each record. Each pulse train was generated with one to ten pulses (amplitude of  $\frac{1}{4}$  to  $\frac{1}{2}$  the amplitude of the signal). Each pulse was 0.005 to 0.1 seconds in length and separated by  $\Delta t = 0.01$  to 0.4 seconds. The parameters were all randomized according to a uniform distribution. The pulse trains were adjusted to achieve SNR values of 0 - 7dB.

To include SAT, all of the data points in a record above a threshold value were reset to the threshold. The threshold was adjusted to achieve percent saturation levels of 10 – 90%, which translated to SNR values of 0.97 - 13.81dB. SNR for SAT was measured as a ratio between the total power in the signal and the power in the signal above the threshold (since the negative of this portion of the signal represents additive noise).

### 3.3 Configuration testing

Fifty one-second signals were simulated using the same method and parameter settings to generate the clean test signals as delineated in Section 3.1-2. Copies of the fifty test signals were then made so that contaminated test signals could be generated by adding simulated noise. For example, Figure 6 illustrates the copying/contaminating process for the signals contaminated with PL. The 50 clean signals were copied nine times producing nine sets of 50 clean signals. Then, each set was contaminated with a different SNR amount of simulated PL producing nine sets of 50 signals with contamination ranging from from PL0 to PL7. This process was repeated for motion artifact and saturation.

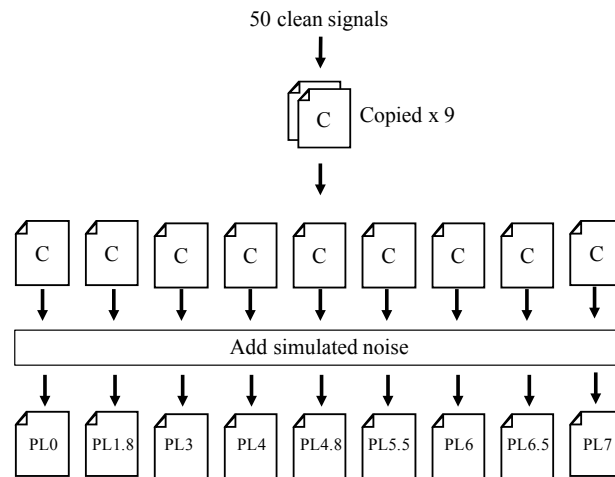


**Figure 6: Process to add simulated noise to clean test signals**

Noise factors were first found with PANDA for the baseline signals. Then, the noise factors for each of the test signals, at varying SNRs, were found.

Testing was completed for all three combinations of feature sets for baseline data set sizes  $N = 300, 600, 1000$ , for bin numbers  $L = 5, 8, 15$ , and for all three noise types (power line interference, saturation, and motion artifact), at all noise amounts.

### **3.4 Results of investigation**

Figure 7 depicts results from one of the configuration set-ups: feature set combination 3 (all features),  $N = 600$  baseline signals,  $L = 5$  bins, with noisy test data PL0 to PL7. The plot on the left shows the noise factors for each of the baseline signals along with the mean across these noise factors and a boundary set at three standard deviations. The plot on the right shows the noise factors from each of the 50 test signals, which were contaminated with increasing SNRs of power line interference. This example demonstrates that PANDA was able to define a boundary between clean and noisy data, capturing 100% of the noisy test samples for SNRs less than 5dB. As the SNR improves, PANDA has more trouble distinguishing between clean and noisy signals, but still managed to capture more than 80% of the noisy signals at SNRs as high as 7dB.

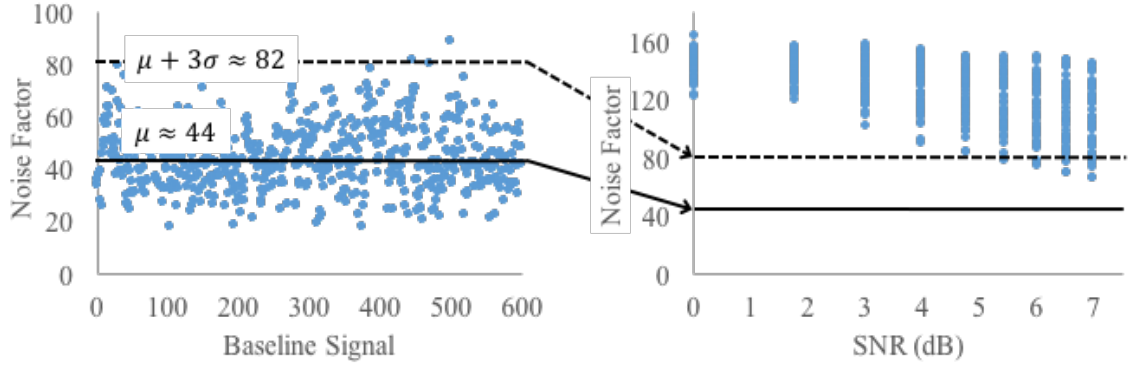


**Figure 7: Noise Factors of baseline signals and signals PL0 to PL7 (feature set combination 3,  $N = 600$ ,  $L = 5$ ).**

Figure 8 depicts % sensitivity (probability of PANDA in detecting noisy signals) for all of the testing for power line interference. %Sensitivity (S) was used as the performance metric defined as:

$$100\% * (\# \text{ noisy signals above boundary}) / \text{Total Number of noisy signals}$$

Each of the nine subplots shows the results of testing for the indicated combination of baseline set size and number of bins. The blue, green, and red lines indicate the set of features used in testing; the sets are [MA, ZC, SSC, WL], [EN, 60Hz, 180Hz, SM], and all features, respectively. The green line in the  $N = 600$ ,  $L = 5$  subplot represents the results shown in Figure 7. Each point indicates the % sensitivity at the indicated SNR value.

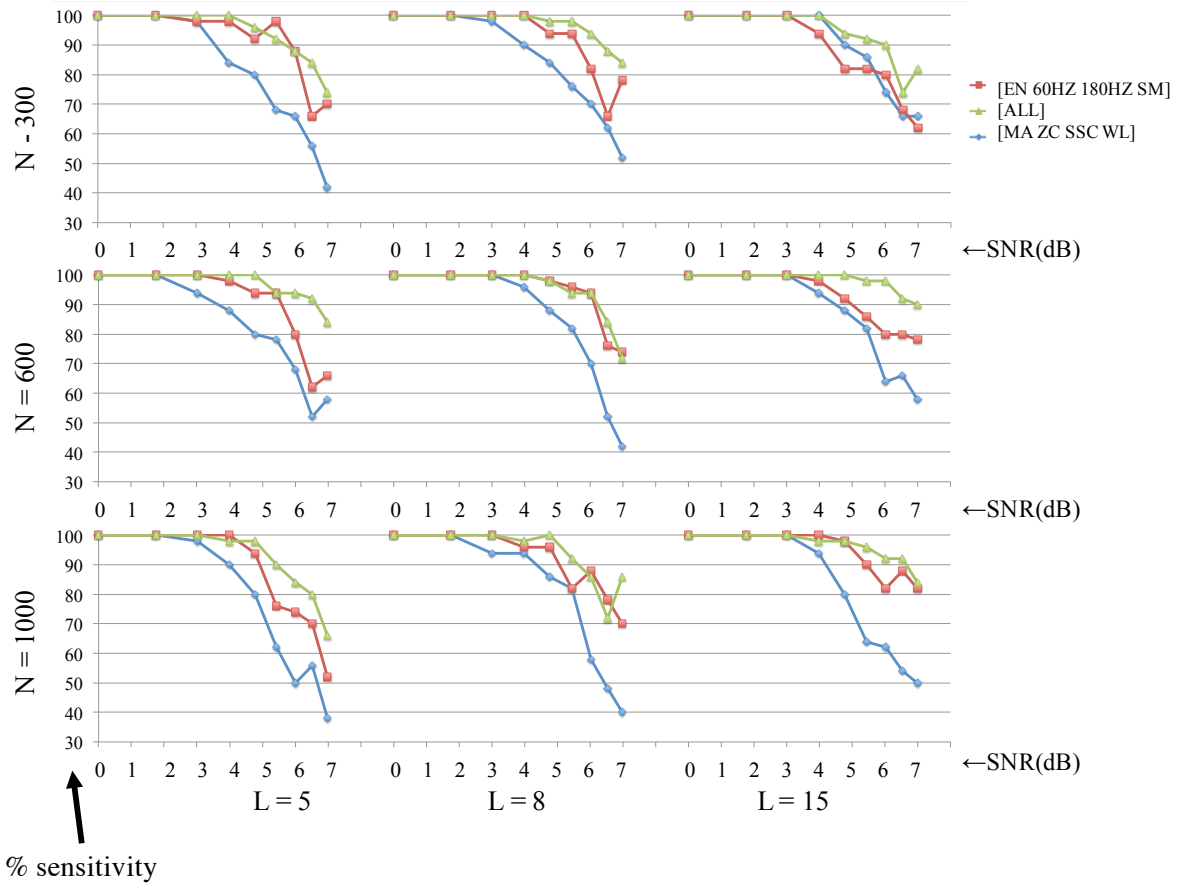


**Figure 8: Testing results of initial investigation of PANDA with simulated baseline and test data. Noise type is power line interference.**

Figure 8 demonstrates that there is very little difference between most configurations for identifying power line interference. However, a moderate improvement in sensitivity was indicated whenever the full set of features was used (green line), and performance stayed better for higher SNR when  $N=600$  records and  $L = 5$  bins or 15 bins of equal width were used.

Similar investigations were completed for motion artifact and saturation. PANDA was also able to successfully identify the contamination for these noise types with 100% sensitivity at high levels of contamination (low SNR), with reduced sensitivity as the noise level decreased. Both were also better identified when all features were used.

There were only marginal improvements with the other factors. For motion artifact, performance improved slightly when  $N=1000$  records, and  $L=15$  bins of equal width. For saturation, marginal improvements were noted when  $N=600$  records and the number of bins had no effect.

### **3.5 Proposed configuration**

Based on the results of this exploration, the factor that was most influential on performance was feature set. It was evident that a full set of features (all eight) was best suited to detect noisy signals using PANDA for all three types of noise. It was less clear how many baseline records and bins were required to provide the best performance. Neither factor had much influence, but for power line interference and motion artifact,  $N=600$  records worked best, and for motion artifact,  $N=1000$  records worked best. For power line interference  $L=5$  bins or 15 bins worked equally well; for saturation it didn't matter, but for motion artifact,  $L=15$  bins worked best. Since performance was marginally affected by these factors and no best choice was obvious, we also decided to consider computational time when choosing a final configuration, favoring lower record and bin counts. As a result, the final configuration for PANDA was set to i) a full feature set, ii)  $N = 600$  baseline records, iii)  $L = 5$  bins with equal bin widths, and iv) boundary setting of three standard deviations from the mean.

## **Chapter 4. PANDA in use with simulated baseline data**

The intent of this work was to use PANDA to detect contamination in recorded signals using simulated signals as the baseline data. It is difficult to record clean data; therefore, using simulated data as the clean baseline data set would improve efficiency of this noise detection approach. The work reported in Chapter 3 indicated that PANDA worked well to detect noisy test data when it was simulated. The purpose of this investigation was to determine how well PANDA worked on recorded test data when the baseline data was simulated.

### **4.1 Methods**

#### **4.1-1 Recorded data collection**

Surface electromyography data was recorded from the biceps brachii muscle using the Delsys TRIGNO™ Wireless System (CMRR>80dB, Bandwidth: 20-450 Hz, resolution  $\approx 168\text{nV/bit}$ , Gain $\approx 300\text{V/V}$ ). Recordings were collected at a sampling rate of 2000 Hz from 12 adults. The electrodes were placed on the muscle according to the guidelines established by SENIAM [11]. The participants were chosen to establish a broad selection of data; seven participants were male and 5 were female, and age ranged from 23 to 74 years (mean  $\approx 38.83$ ; standard deviation  $\approx 18.75$ ). Prior to collection, the skin where the electrodes were placed was cleaned with an alcohol pad and rubbed with conductive electrode gel. The electrodes were placed over the muscle belly and in parallel with the muscle fibers.

Experimentation complied with Tri-Council Ethical Regulations and all participants gave informed consent. Each participant was placed in a chair and asked to find a

comfortable position. The humeral portion of the arm was held close to the side of the body with the elbow bent at a 90-degree angle. In this position, each participant was asked to find a comfortable way to further contract their biceps brachii. Some chose to clench a fist or squeeze a piece of provided foam.

The participant was asked to elicit a strong biceps brachii contraction for a short period of time (approximately eight seconds). Using the data collection tool shown in Figure 9, a line indicating 50% of the strong contraction was displayed in the plot area, which was visible to the participant. The participant was then asked to hold a static contraction for two minutes at the indicated 50% target; this satisfied the aim of producing a medium-intensity contraction. Real-time monitoring of the contraction level was visible on the screen to guide each participant. Verbal encouragement was also provided for the participants to maintain the contraction.

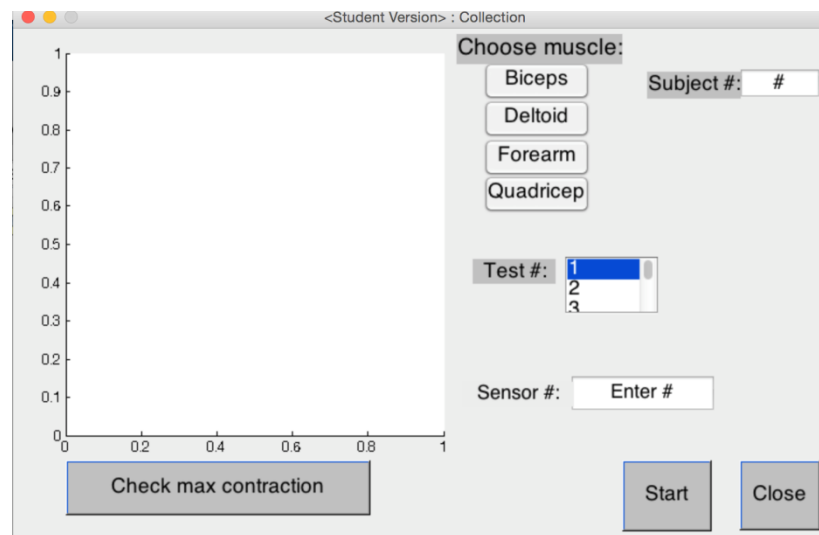


**Figure 9: Data collection tool**

In post processing, twenty-five 1-second records were extracted for analysis from the long contraction record from the 5-second mark until the 30-second mark. All records

were visually inspected for obvious contaminants. Then, two techniques were completed to verify the cleanliness of the collected data. First, spectral interpolation was conducted at 60 Hz and 180 Hz according to [12] to remove power line interference and its harmonics. Then, an evaluation was carried out using a CleanEMG tool. The tool tested the data for power line interference according to methods described in [12] and [16]; saturation, quantization noise and motion artifact according to methods found in [13]. Any data not meeting the cleanliness criterion was thrown out.

In total, 300 records were processed (25 1-sec segments from 12 participants) and 10 records were discarded yielding a clean test data set of  $N=290$  records.

#### **4.1-2 Record contamination**

Contaminated test data sets were generated using copies of the one-second test records, similar to the contamination process described in Chapter 3. For each noise type, a set of copies of the clean records were artificially contaminated at increasing levels. Figure 10 illustrates examples of each of the noise signals, along with the resulting contaminated records.

Three types of noise were tested: power line interference (PL); motion artifact (MA); and saturation (SAT). The copies of the records were contaminated with the noise using the same process as described in Section 3.2. For PL, the amplitude of the sinusoid was adjusted to achieve SNR values of -20 to 30dB. For MA, duty cycles of the pulse trains were adjusted to achieve SNR values of -20 to 30dB. And finally for SAT, the threshold was adjusted to achieve SNR values of 0.07 to 33.7dB. These ranges were chosen to

explore the behavior of the algorithm for very small to very large amounts of additive noise.

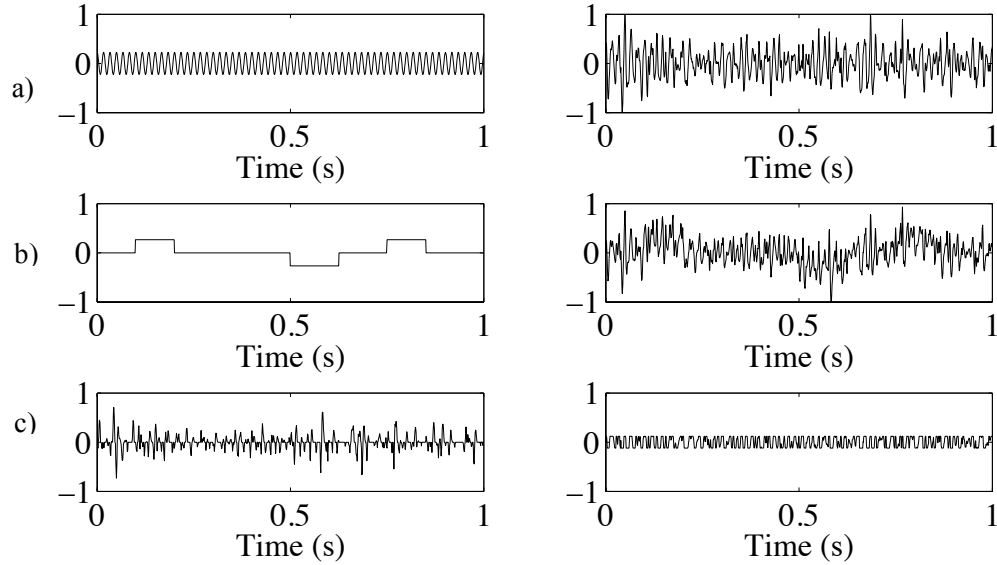


**Figure 10: Process used to add contaminants a) Power line interference SNR 3dB b) Motion artifact SNR 3dB c) Saturation 3dB**

### 4.1-3 Testing method

PANDA was configured as described in Chapter 3 with i) the full feature set, ii)  $N = 600$  simulated baseline signals (using the process as described in Section 3.1-2), iii)  $L = 5$  bins, and iv) a boundary setting of three standard deviations from the mean. A noise factor for each of the baseline signals was generated and the mean and boundary settings were established. Then, noise factors for the clean and contaminated test data sets were generated.

## 4.2 Results

Figure 11 - Figure 13 display the results of testing clean recorded data and contaminated recorded data sets at PL7, MA7, and SAT6.7, respectively. In the plots, the blue points

represent  $F_N$  values for the baseline data; the green points represent  $F_N$  values for the clean test signals; and the red points represent  $F_N$  values for the contaminated recorded data. As depicted in the figures, the mean of the baseline data  $F_N$  values was 44.6 and the boundary ( $\mu + 3\sigma$ ) was 81.7. For PL7 and MA7, PANDA identified 100% of the contaminated signals as being contaminated and for SAT6.7, PANDA identified 98% of the contaminated signals as being contaminated. However, the algorithm also identified 96.2% ( $290 - 11 = 279$ ) of the known-to-be-clean test signals as being contaminated.

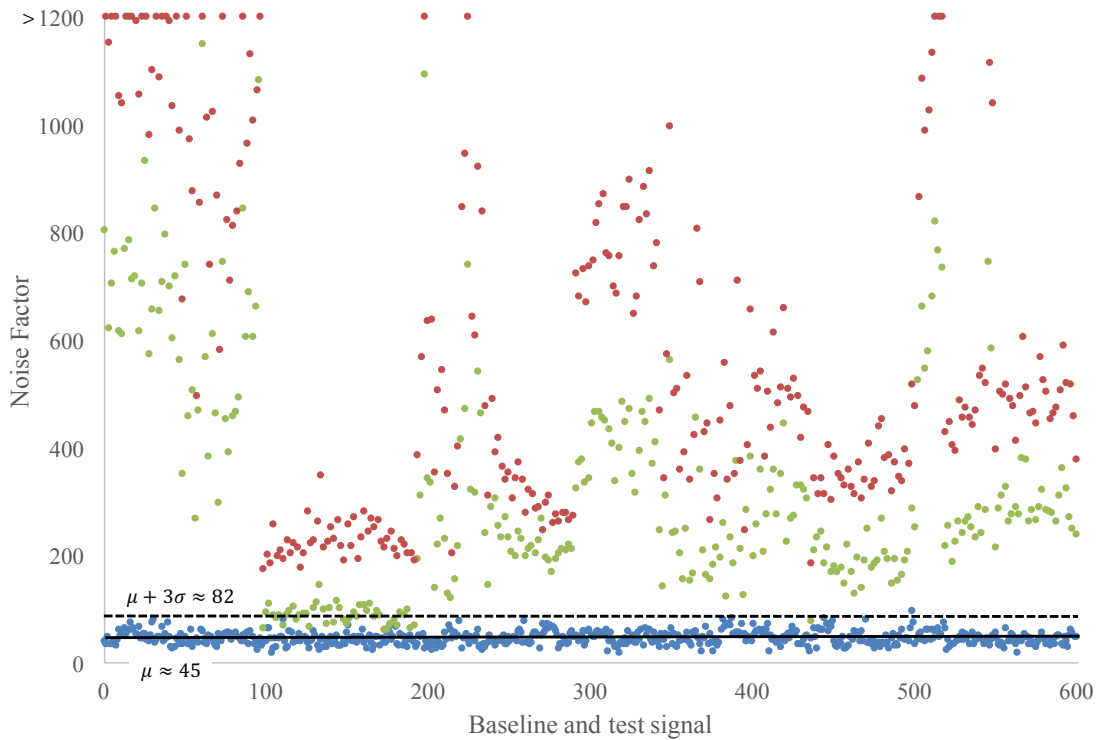


**Figure 11: Simulated baseline data testing with recorded test data. Simulated  $F_N$  in blue, clean recorded  $F_N$  in green, and contaminated (PL7) recorded  $F_N$  in red.**

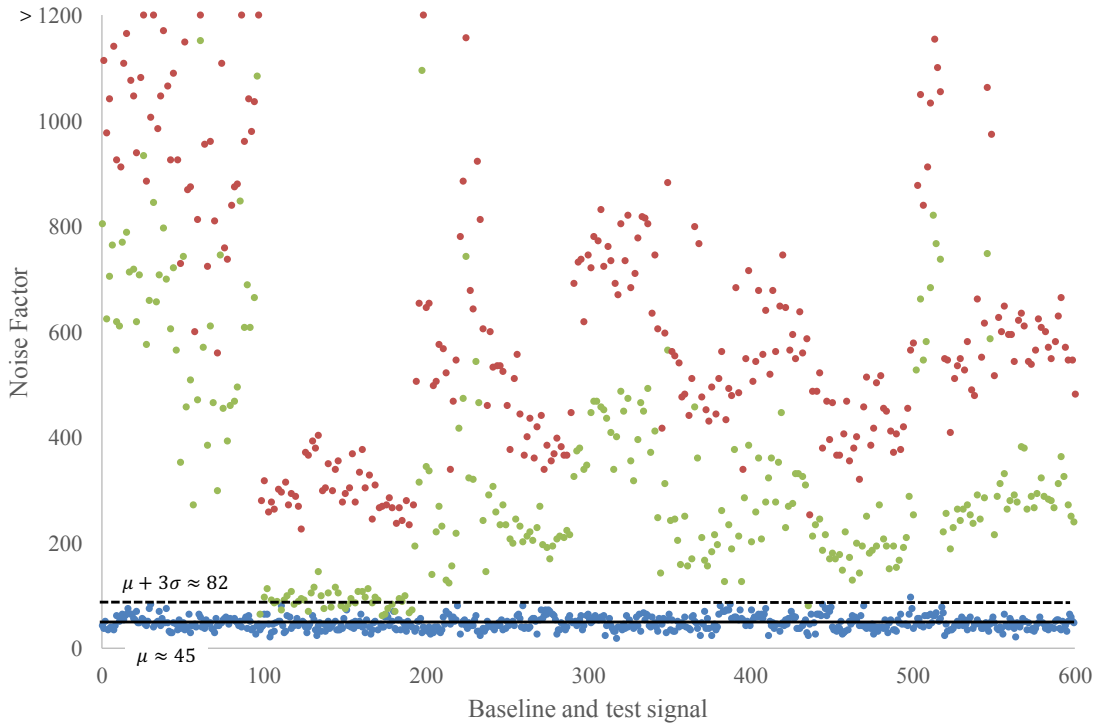


**Figure 12: Simulated baseline data with recorded test data Simulated  $F_N$  in blue, recorded  $F_N$  in green, and contaminated (MA7) recorded  $F_N$  in red.**

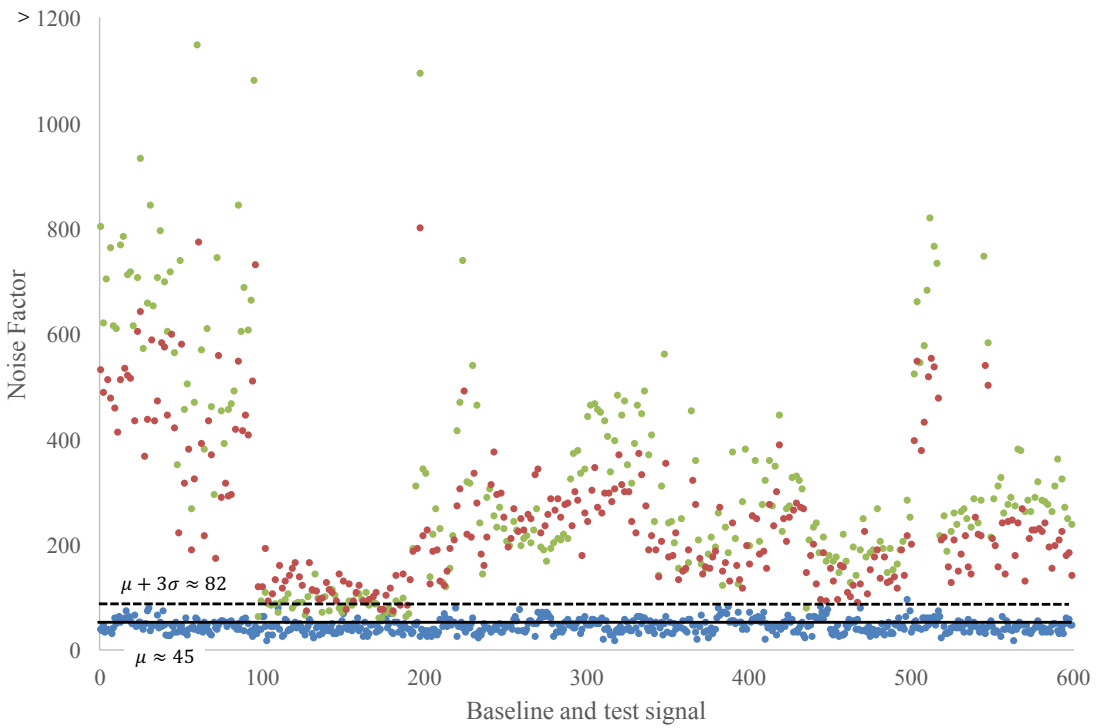


**Figure 13: Simulated baseline data with recorded test data. Simulated  $F_N$  in blue, recoded  $F_N$  in green, and contaminated (SAT6.7) recorded  $F_N$  in red.**

For each contaminant type, PANDA's ability to detect contaminated signals degraded as the SNR increased; however, at the highest SNR for each contamination, PANDA still had high detection rates: 95.9% sensitivity at PL30; 92.1% sensitivity at MA30; and 94.1% sensitivity at SAT33.7. PANDA was able to correctly identify all signals with 100% sensitivity that met the following conditions:

- Contaminated with power line interference having SNRs  $\leq 15\text{dB}$
- Contaminated with motion artifact having SNRs  $\leq 15\text{dB}$
- Contaminated with saturation having SNRs  $\leq 5.22\text{dB}$  (equivalent to 40% saturation)

While PANDA's detection rate of contaminated signals was high, it is important to remember its false positive rate was found to be 96.2%.

### **4.3 Discussion**

In this investigation, simulated data was used as baseline data, while recorded data, collected from the biceps of 12 participants, was used as test data. While PANDA was able to correctly identify noisy test signals as being noisy, it performed poorly in the identification of clean test signals. Moderate improvement could be achieved by moving the boundary up. For example, if the boundary in Figure 11 was moved to  $\sim 168$  (equivalent to the mean plus 10 standard deviations), it would still accurately identify 100% of the contaminated test signals, but would identify only 76% ( $290 - 70 = 220$ ) of the known-to-be-clean test signals as being contaminated instead of 96.2% as with the

previous boundary. Nevertheless, this moderate improvement is not sufficient to make the technique practically viable and the poor results indicate that PANDA is having a hard time comparing between simulated and recorded clean EMG data. PANDA's ability to correctly identify the contaminated signals as noisy is of little value given its inability to correctly identify the clean test data as clean.

Three limitations of this work were identified, which may account for the trouble PANDA had with comparing simulated with recorded data – 1) the lack of baseline noise in the baseline data, 2) the homogeneity of the baseline data, and 3) the potentially poor setup of the physiological parameters of Myosim. Inherent to sEMG signals (even clean sEMG signals) is baseline noise. This noise is very small, but may affect the simulated signal's comparison to the recorded data. Also, the simulated data was modeled using parameter settings that mimic a somewhat homogenous set of data records. While the number of fibers per motor unit and the location of each motor unit varied across data records, the number of motor units firing and the firing statistics were held constant; so was limb size, which constrained the extent to which motor unit location could be varied. The variations in model parameters may not have sufficiently captured variation expected across different participants engaging in a self-directed medium-level contraction. Perhaps the simulated data was too homogenous for PANDA to account for the variation in recordings. Finally, the values of the physiological parameters for the sEMG simulator, Myosim, may not have been set to appropriately model a typical bicep contraction. While we followed parameter settings reported in the literature [26], we recognize the difficulties associated with estimating physiological parameters (like number of motor units firing) that cannot be directly measured. If more

work can be done to overcome these shortcomings and improve the simulated data set so it better represents recorded data, using PANDA with simulated baseline data may still be possible. Alternatively, carefully collected recorded data can be used to establish the baseline data set.

## **Chapter 5. PANDA in use with recorded baseline data**

### **5.1 Purpose**

The purpose of this investigation was to test the performance of PANDA in detecting contaminated sEMG recordings with recorded data employed as both the baseline and test data sets. Our original intentions were to develop PANDA for use with simulated baseline data, but as demonstrated in the work reported in Chapter 4, this approach was not possible. In this investigation, we aimed to see if PANDA worked better with recorded baseline data.

### **5.2 Methods**

#### **5.2-1 Recorded data collection**

Since this investigation required recorded data for both baseline and test data sets, another data collection protocol was employed to add to the original recorded data set of N=290 1-sec records. During the same data collection session as outlined in Chapter 4, each participant was asked to perform another set of contractions with the electrodes and instrumentation still in place from the first collection. As in the first protocol, each participant was first asked to elicit a strong biceps brachii contraction for a short period of time. They were then asked to hold a static contraction at a target of 50% of their strong contraction (indicated on a computer screen). However, during this second protocol, instead of holding one long contraction for 2 minutes (as in the first protocol), participants held the contraction for 10 seconds and followed it with a five-second rest period and repeated this contraction/rest combination five times.

In post processing, ten 1-sec records were extracted for analysis from each of the short contraction records. All records were then processed as described in Chapter 4 to

remove any contaminated records. In total, 600 records were processed (5 sets of 10 1-sec segments from 12 participants) and 13 records were discarded yielding a new data set consisting of  $N = 587$  records. Adding this to the 290 records from the first data collection yielded a summative total of  $N = 877$  records.

### **5.2-2 Record contamination**

Record contamination followed the same process as detailed in Chapter 4 yielding the following: for power line interference, the amplitude of the sinusoid was adjusted to achieve SNR values of -20 to 30dB; for motion artifact, duty cycles of the pulse trains were adjusted to achieve SNR values of -20 to 30dB; and finally for saturation, the threshold was adjusted to achieve SNR values of 0.07 to 33.7dB.

In addition, two contamination combinations were included by blending two or more types of noise. A blended contamination of power line interference and motion artifact was included by adding equal amounts of the two noise types, yielding blended SNR values of -40.2 to 27.0dB. A blended contamination was also included by simultaneously adding all three noise types; equal amounts of power line interference and motion artifact were added along with percentages of saturation of either 0.5% or 1% to yield blended SNR values from -40.2 to 24.8dB. Following the naming convention established in Chapter 3, the first blended noise type is referred to as PL-MAx and the second blend is referred to as PL-MA-SATx, where x is the SNR in dB.

Multiple copies of the clean test data set were made for each of the five noise types (three separate and two blended). For each, the multiple copies were contaminated with varying amounts of noise to achieve the stated SNR values as depicted in Figure 6.

Figure 14 illustrates examples of each of the noise signals, along with the resulting contaminated records.

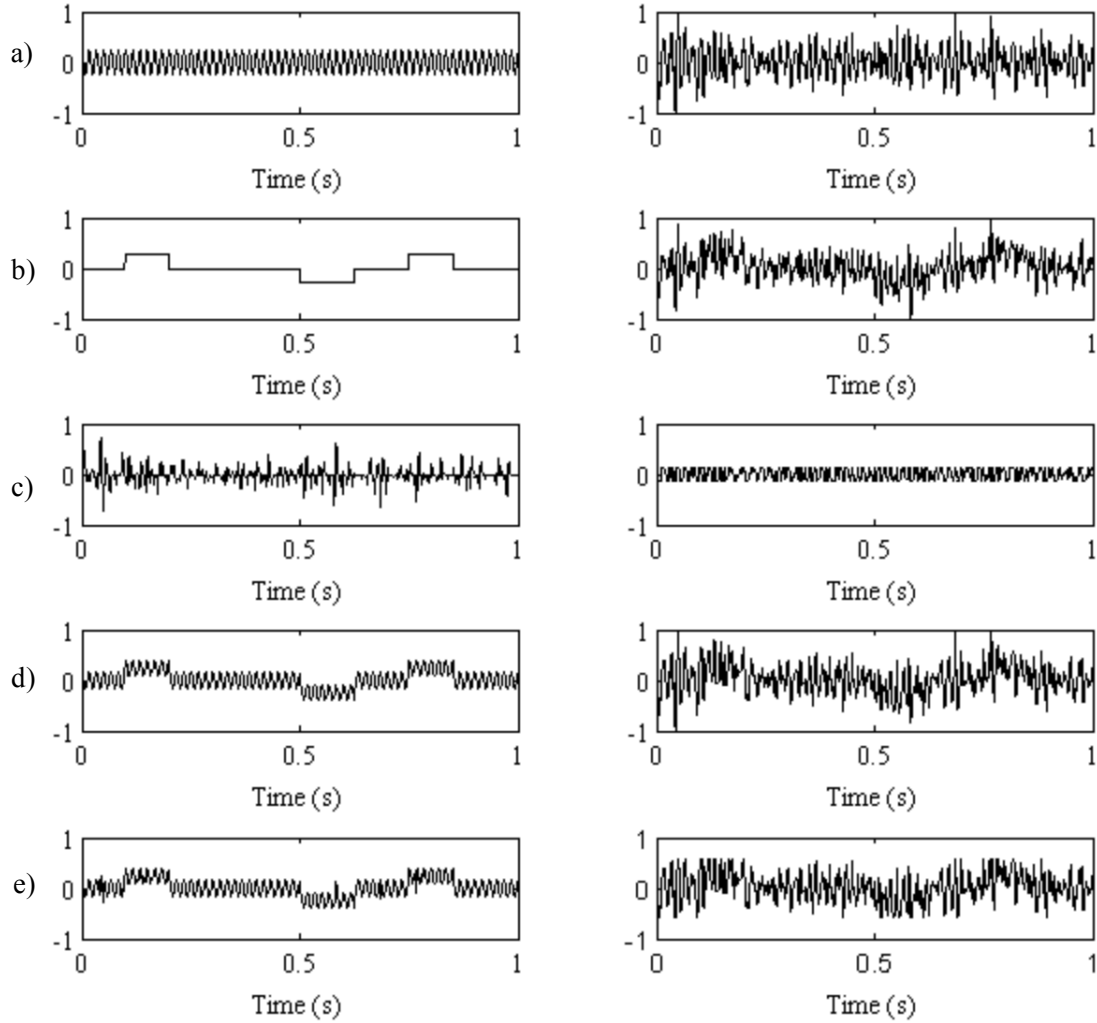


**Figure 14:** Process used to add contaminants a) PL3 b) MA3 c) SAT3 d) PL-MA4.3, which is equivalent to the addition of PL6 and MA6 e) PL-MA-SAT4.3, which is equivalent to the addition of PL6, MA6, and SAT18.8.

### 5.2-3 Testing method

The configuration of PANDA was similar to the configuration in Chapter 4, but in an effort to try and improve on its performance, adjustments were made. First, normality was introduced as a ninth feature. By the time this portion of the work was conducted,

other researchers involved in the CleanEMG project had used measures of normality in their work [18], so this feature was added to the PANDA feature set. Normality in the amplitude distribution of an SEMG record is represented by this feature, which is calculated by summing the first and tenth bins of the 10-bin histogram of the record. The feature was added to help PANDA in detecting saturation.

The second adjustment involved the number of baseline records used. Since we were using recorded data, we were limited in choice by the number of recorded data sets available. Fortunately, we had sufficient data to include more than N=600 records (the value proposed in the original configuration). Using 12-fold cross validation provided for N=825 records for use in the baseline data is described below.

N-fold stratified cross validation is a technique used in machine learning to evaluate a learning system by withholding an N<sup>th</sup> of data from training, and using that data to validate the performance achieved by training on the rest. For this investigation 12 ‘folds’ of data were available (one data set containing 75 records from each of 12 participants). Eleven folds were used as baseline data (yielding 75 x 11 = 825 records) and performance was measured using the remaining fold as test data. %Sensitivity (S) was used as the performance metric defined as:

$$100\% * (\# \text{ noisy signals above boundary}) / \text{Total Number of noisy signals}$$

The resolution of this metric depends on the number of signals tested. Since the 12<sup>th</sup> fold included 75 records, the resolution was  $R=1/75$  or about 1%. To evaluate precision, the validation process was repeated 12 times, using a different participant data set as the hold-out set for each repetition. Further to the contaminated test data set

organization as described in the previous section, the method to obtain each copied data set for each noise type and combination was replicated 12 times for each of the participant testing sets.

The third adjustment involved using receiver operating characteristic (ROC) analysis to determine a suitable choice for 1) the number of bins used and 2) the boundary setting. ROC analysis delineates the performance of a binary system as a parameter or threshold is varied.

This analysis was performed to identify the most suitable number of bins to use per noise type and the most suitable boundary setting per noise type. A weakness of using the three standard deviation boundary is that clean data could be thrown away inadvertently; therefore, the ROC testing was done to establish a threshold that would keep the most number of clean records while throwing away an acceptable number of contaminated records.

Figure 15 illustrates the probability density functions (PDF) model of a clean and contaminated data set; these correspond to the ROC testing criteria illustrated in Figure 16. The vertical dashed line indicates the threshold. The PDF colored in blue indicates the probability of a clean  $F_N$  and the yellow PDF indicates the probability of a noisy  $F_N$ . The portion of the clean (blue) PDF to the other side of the threshold (indicated in green), specifies the probability that PANDA is incorrectly identifying those clean signals as being contaminated; this is referred to as the probability of a false alarm or a False Positive. Similarly, the portion of the yellow PDF to the other side of the threshold (orange), specifies the probability that PANDA is identifying noisy signals as being

clean; this is referred to as the probability of a miss or a False Negative. The threshold could be moved to the left in order to decrease the probability of a miss, but this would increase the probability of false alarm and vice versa. Modification of the number of bins will change the distribution of the PDFs, which has the potential to increase or decrease the probabilities of a false alarm or a miss.

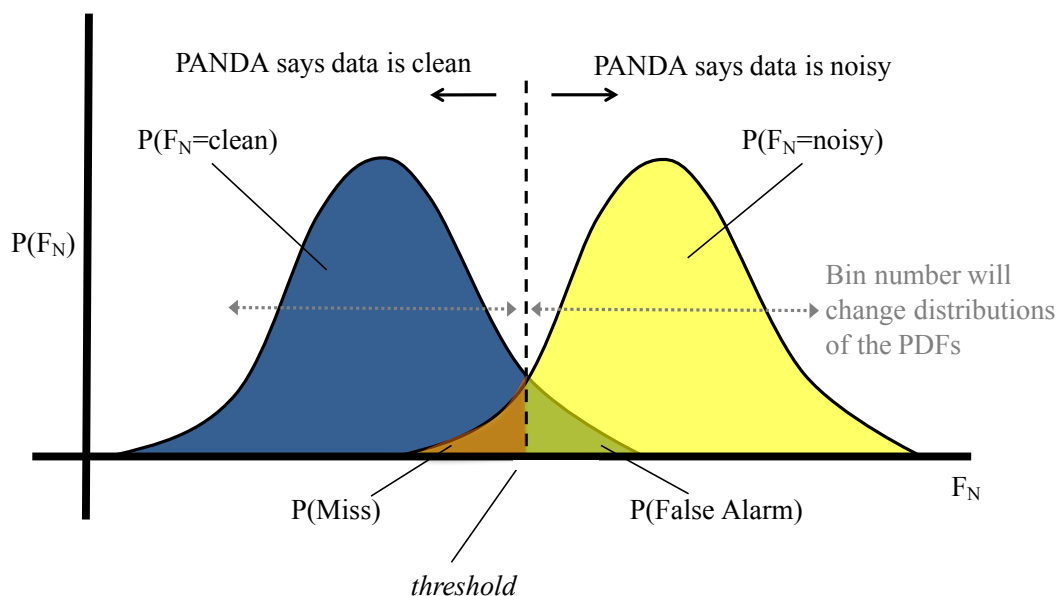


Figure 15: Probability curve

|                     |       |                          |                                 |
|---------------------|-------|--------------------------|---------------------------------|
|                     |       | <b>PANDA says:</b>       |                                 |
|                     |       | Clean                    | Noisy                           |
| <b>The data is:</b> | Clean | True Negative            | False Positive<br>(False Alarm) |
|                     | Noisy | False Negative<br>(Miss) | True Positive                   |

Figure 16: ROC testing criteria

ROC testing compares the False Positive rate to the True Positive rate. A visualization of the two is created and the parameter in question (in this case – bin number or boundary setting) can be assigned that best minimizes the rate of false negatives and maximizes the rate of true positives.

In this testing, the number of bins tested were 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, and 20. The boundary setting was tested by modifying the multiplier of the  $F_N$  standard deviation. Multipliers of 1, 1.5, 2, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4, and 4.5 were tested.

Figure 17 gives the ROCs for the comparison of the clean test set and the test set contaminated with power line interference.

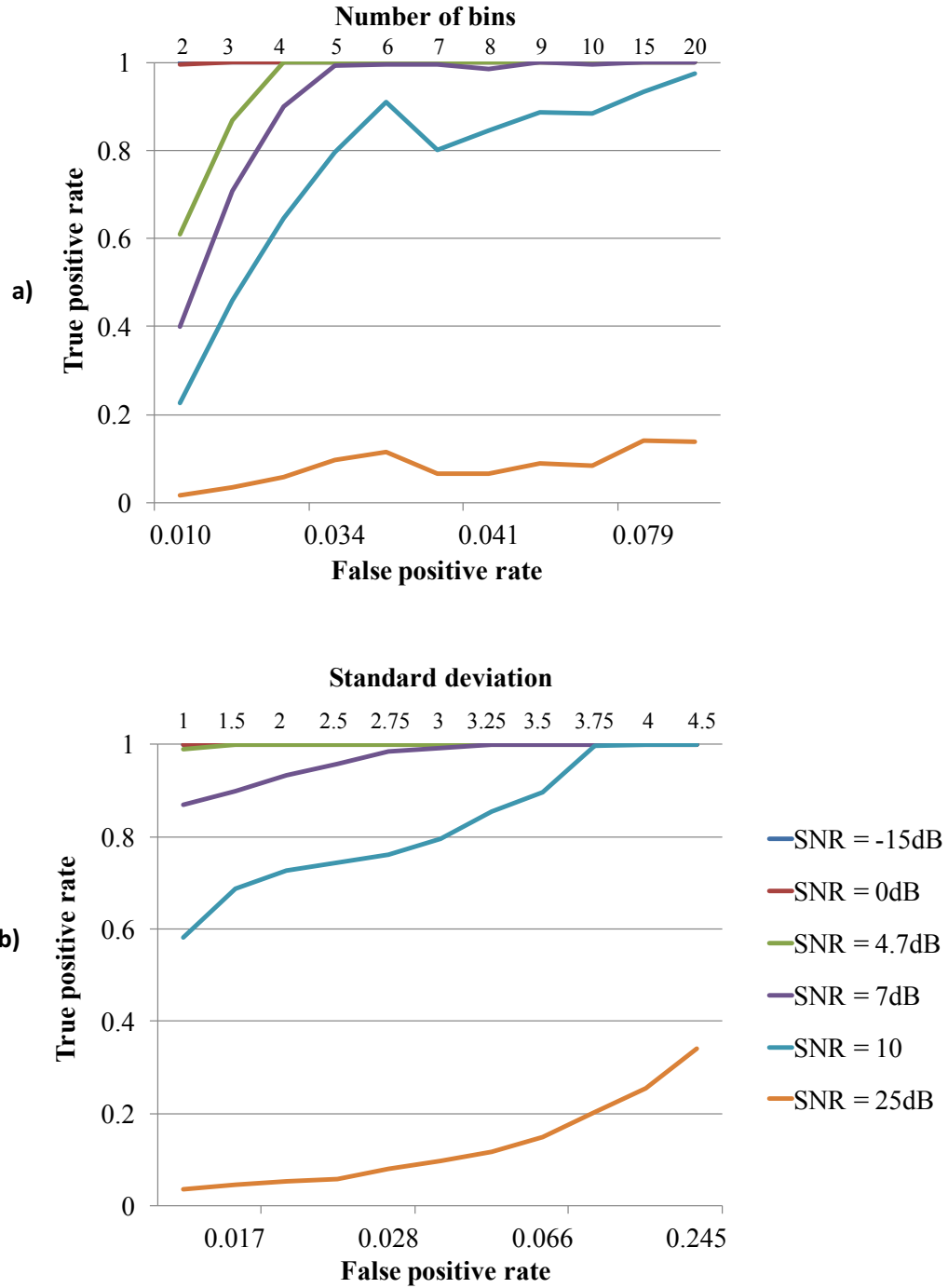


**Figure 17: ROC investigation of a) bin number (for a SD multiplier of 3) and b) standard deviation boundary multiplier (for a bin number of 5) for the addition of power line interference to sEMG at varying SNRs**

ROC testing revealed that the most suitable number of bins for power line interference and motion artifact was five. For saturation, the suitable number of bins was found to be six. The standard deviation multiplier setting of three was found to be the most suitable for all three types of noise. Based on these findings, settings of five and three were used for the bin and multiplier, respectively, for the combined noise tests.

Once PANDA was configured, each of the 12 test-baseline data folds were processed. In each run, noise factors for each of the baseline signals were generated and the mean and boundary settings were established. Then, noise factors for the clean and contaminated test data sets were generated and %sensitivities were assessed.

### 5.3 Results

Figure 18 illustrates PANDA's ability to identify clean signals as clean. For each of the 12 test-baseline data folds, the accuracy of PANDA's ability to identify known-to-be clean data as clean (falling below the threshold), was calculated. The data points in Figure 18 indicate the % accuracy for each fold. The solid black line shows the average accuracy for the 12 tests (90.4%). It is evident that the 6<sup>th</sup> data point is significantly lower than the other points (major outlier according to the Schematic Boxplot Outlier Identification Method [27]). This point represents the test for the 6<sup>th</sup> participant. The dashed black line represents the accuracy of the all tests, except the 6<sup>th</sup> (95.8%) to illustrate the accuracy with the exclusion of an outlier. The probability of a false positive or false alarm is therefore  $P(FA) = 9.6\%$  (1-90.4%).

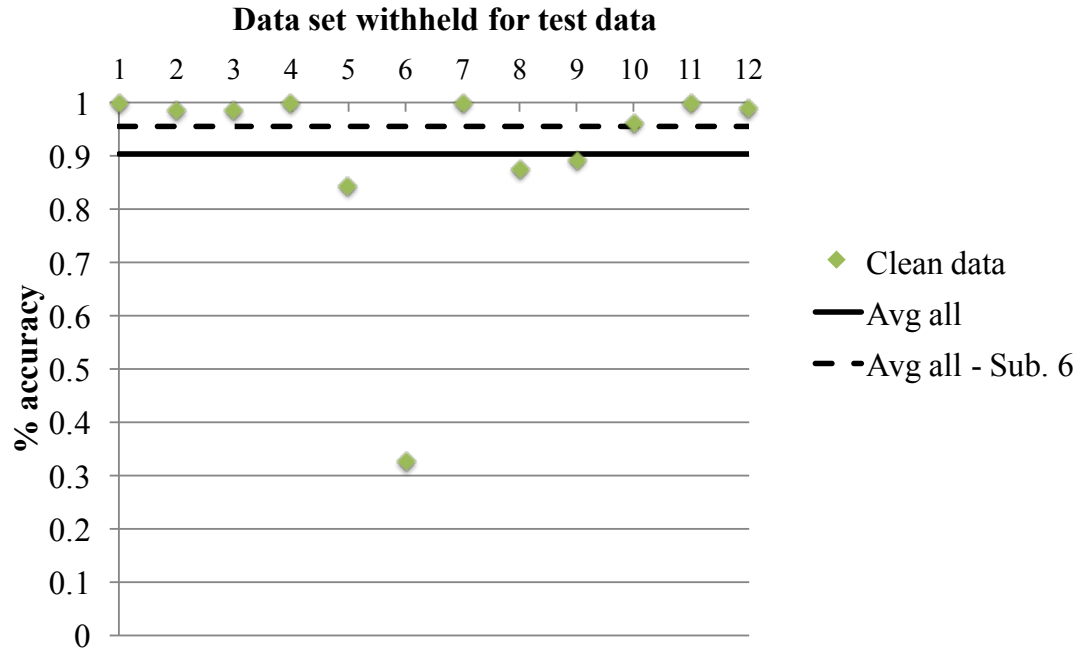


**Figure 18: Sensitivity of clean data detection in PANDA**

Figure 19 - Figure 23 depict the functionality of PANDA in detecting different types of noise at varying SNR amounts. In each figure, part a) illustrates the Algorithm's % sensitivity in detecting contaminated records as a function of SNR. This metric is depicted in the figures as the fraction of noisy records classified as noisy. The data points represent average % sensitivity across all 12 validation folds and the error bars indicate one standard deviation. A transition point, where the sensitivity rate begins to decline, is visible on all plots. In each figure, part b) displays a time domain segment of a clean sEMG record (in black) superimposed on that same record artificially contaminated with noise (in red). Part c) displays the power spectrums of the same records.

As depicted in the figures, PANDA was able to successfully identify all signals with low SNR values (i.e. high amounts of noise) with 100% sensitivity. The transition point for

power line interference is approximately 7dB as indicated in Figure 19. For motion artifact it is approximately 3dB as indicated in Figure 20, and for saturation it is approximately 17dB as indicated in Figure 21. Note that 17dB is equivalent to 2% saturation, meaning that PANDA was able to successfully identify a record of 1000 data points which had 2%, or 20, of its points saturated.

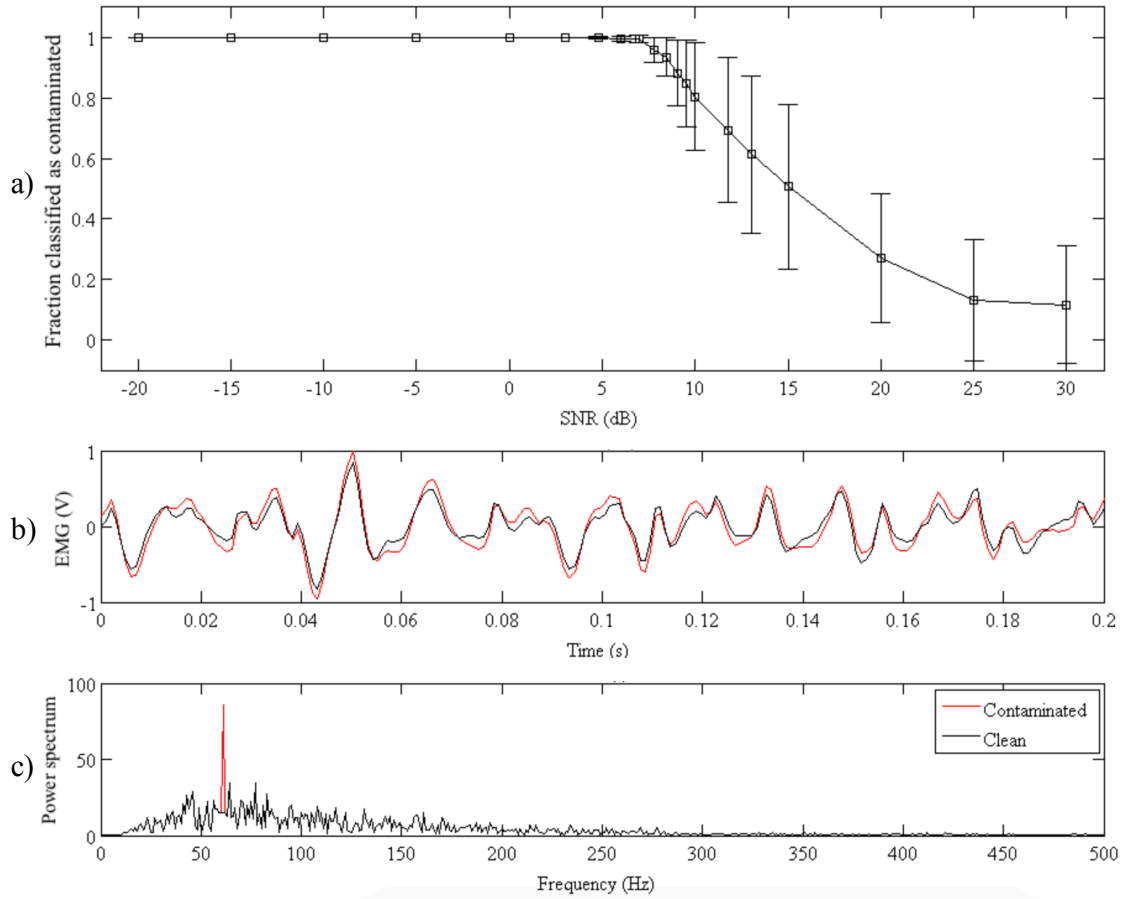


**Figure 19: a) Classification of sEMG contaminated with power line interference (Recall  $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with PL7 (red). c) Power spectrums of clean and contaminated records.**

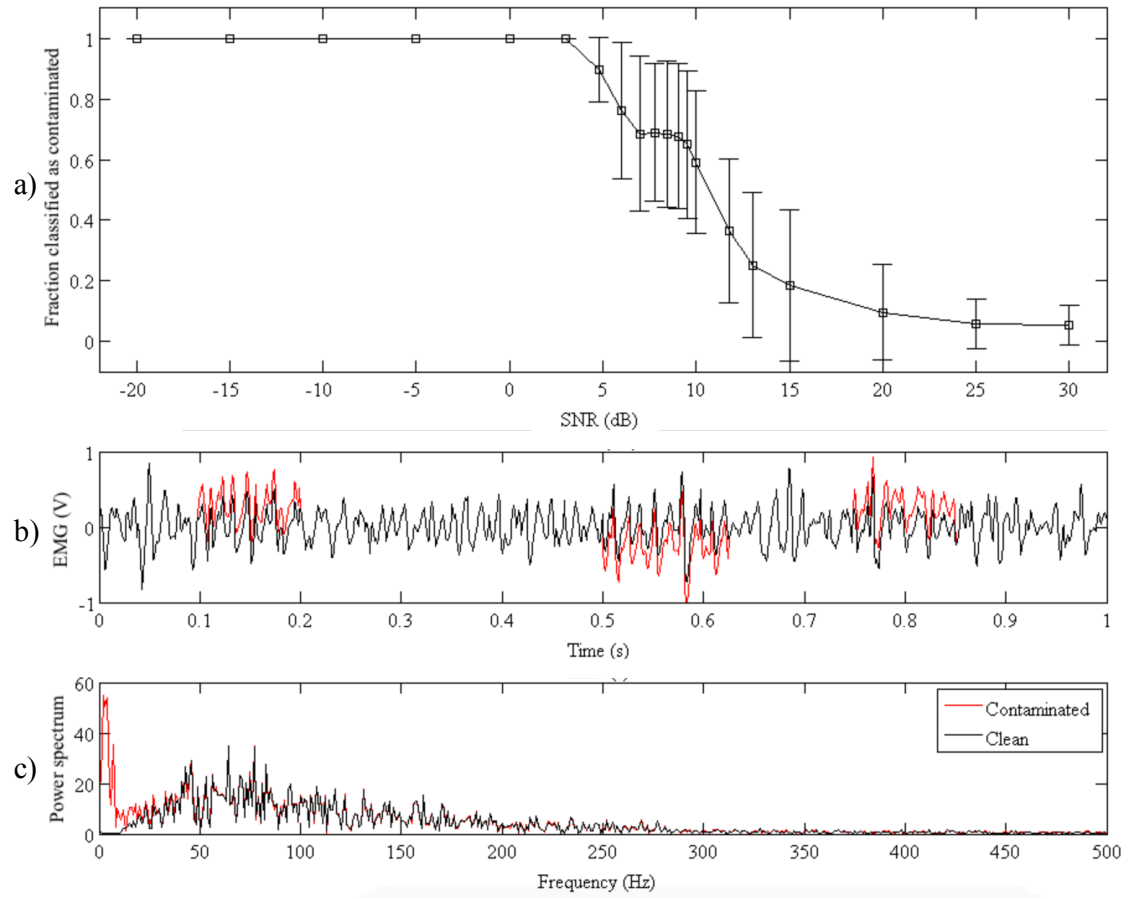


**Figure 20: a) Classification of sEMG contaminated with motion artifact (Recall  $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with MA7 (red). c) Power spectrums of clean and contaminated records.**

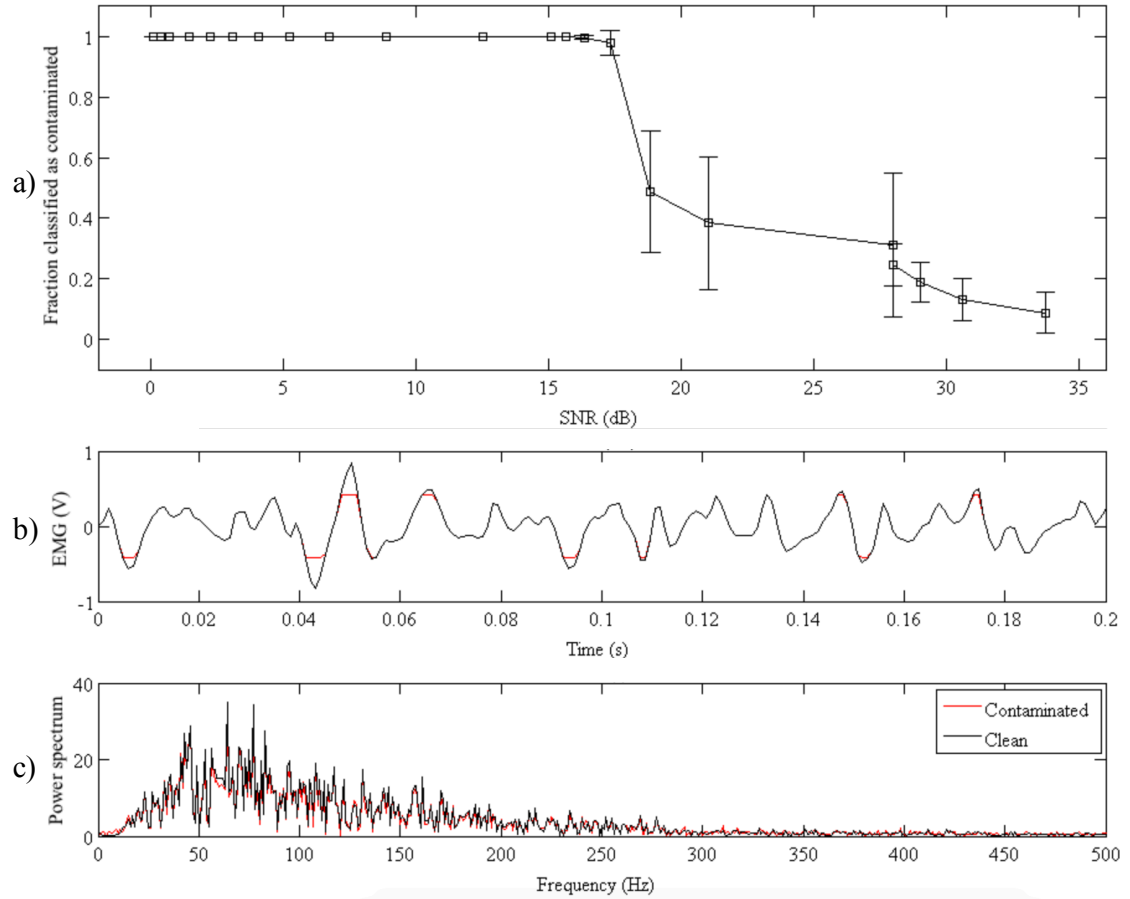


**Figure 21: a) Classification of sEMG contaminated with saturation (Recall  $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with B (red). c) Power spectra of clean and contaminated records.**

Figure 22 and Figure 23 depict the same information as the previous figures but for the blended noise types. To determine the SNR for the blended noise combinations, the power of the clean signal was divided by the power of the noise. The noise was found by subtracting the clean signal from the contaminated signal. Figure 22 indicates a transition point occurring around 7dB measured as a combination of power line interference and motion artifact, which is equivalent to 10dB of power line interference and 10dB of motion artifact. Figure 23 indicates a transition point at about 9dB measured as a combination of power line interference, motion artifact and saturation.

The transition point is the same for both plots, one representing 1% saturation (blue) and the other representing 0.5% (black). Saturation was purposefully chosen to be low because of PANDAs known sensitivity to it.

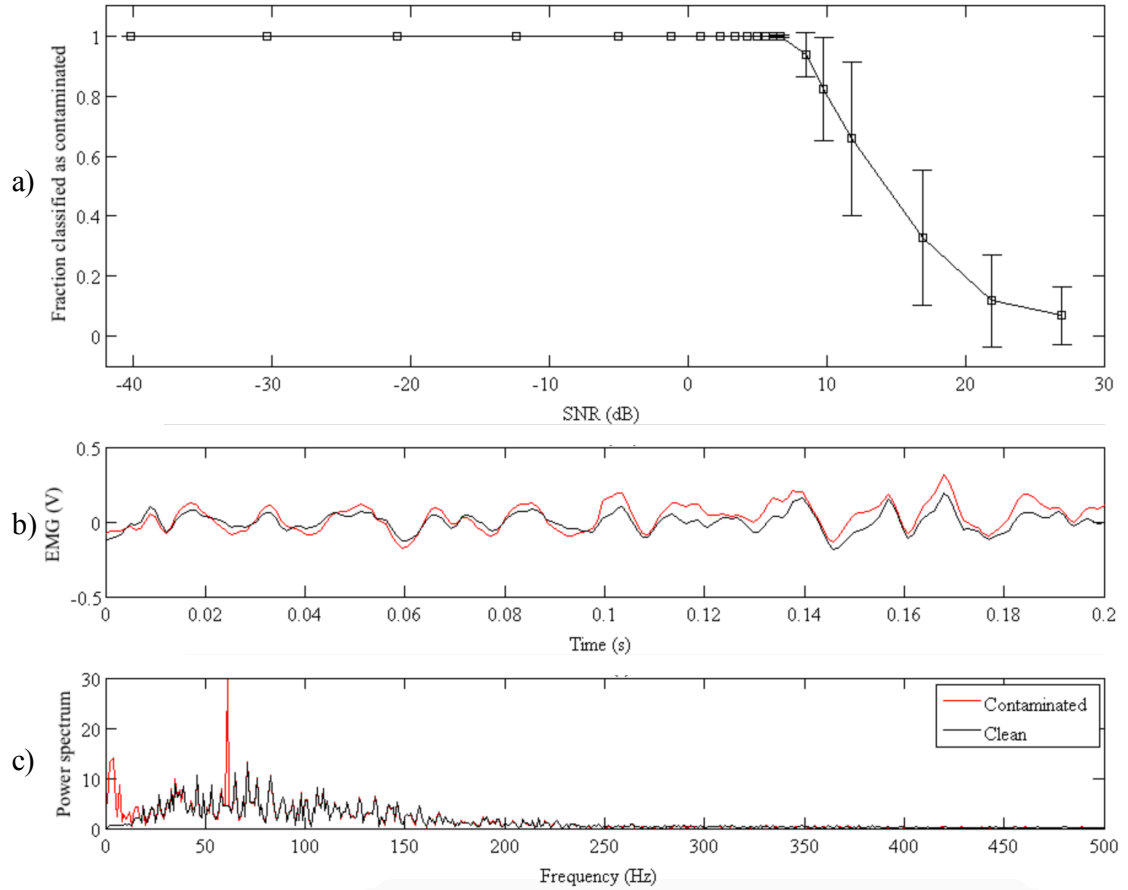


**Figure 22: a) Classification of sEMG contaminated with power line interference and motion artifact (recall  $P(FA) = 9.6\%$ ). b) Time domain of clean sEMG record (black) and sEMG record contaminated with PL3 and MA3 (red), equivalent to PL-MA0.9. c) Power spectra of clean and contaminated records.**

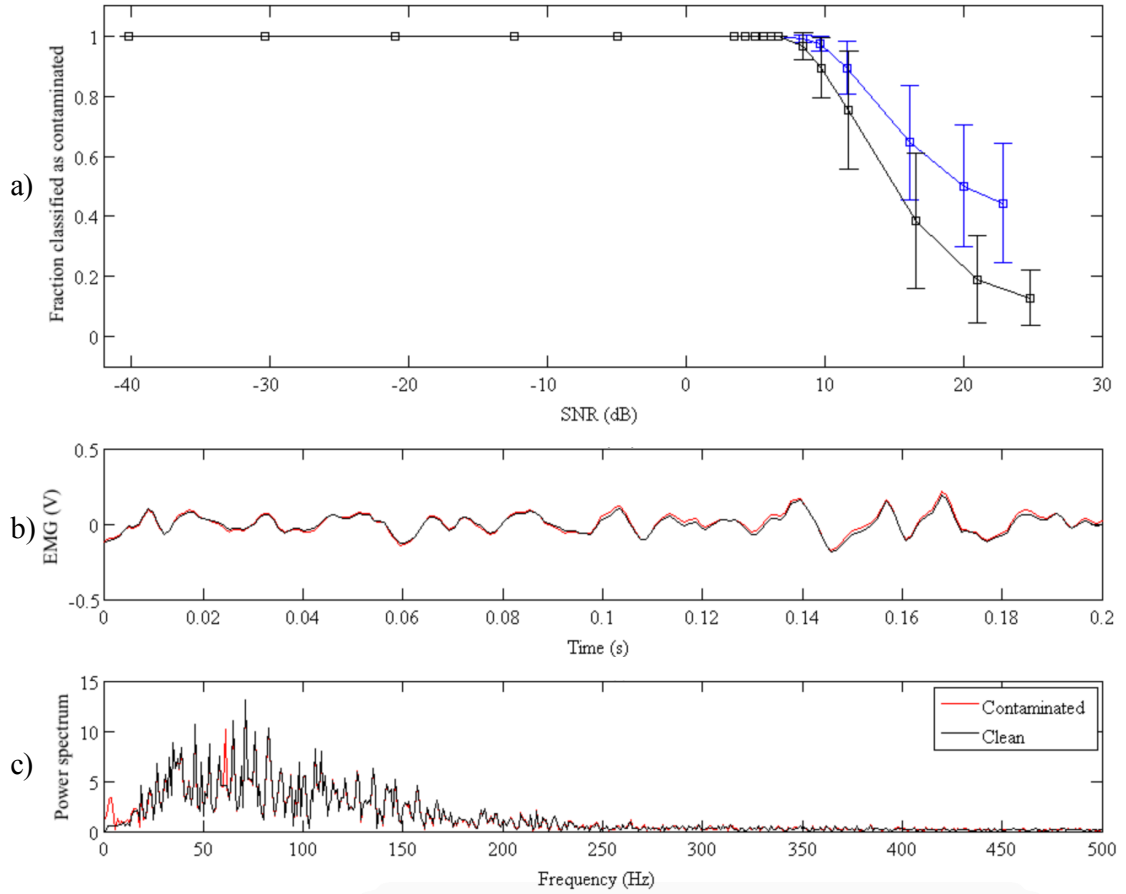


**Figure 23: a) Classification of sEMG contaminated with power line interference, motion artifact, and saturation (recall  $P(FA)=9.6\%$ ). The blue and black represent additions of 1% and 0.5% of saturation, respectively. b) Time domain of clean sEMG record (black) and sEMG record contaminated with PL20, MA20, and SAT18.5 (or 1% saturation) (red), equivalent to PL-MA-SAT16.16. c) Power spectra of clean and contaminated records.**

## 5.4 Discussion

PANDA was tested with recorded data used as both the baseline and test data sets to determine if it could differentiate between clean and contaminated sEMG signals. PANDA was configured using information obtained from testing with simulated data sets, but modified slightly. A ninth feature was used and ROC testing was employed to identify a suitable number of bins and threshold setting. Five contamination types (three separate and two blended) were investigated.

The probability of a false alarm (PANDA identifying a clean signal as unclean) was found to be 9.6%. One of the subjects' data was significantly lower than the rest; in removing this, the P(FA) was lowered to 4.3%. PANDA performed well in identifying contaminated signals as noisy while identifying clean signals as clean. In analyzing the results across SNR values, the transition point (where the sensitivity begins to decline) was evident for each type of noise. PANDA performed best in detecting saturation (100% sensitivity at levels as high as SAT17), moderately well with power line interference (100% sensitivity at levels as high as PL7), and reasonably well with motion artifact (100% sensitivity at levels as high as MA3). Even in the noise combination conditions, PANDA worked at 100% sensitivity at levels as high as SNR=7dB.

As indicated by the ROC analysis, the boundary can be moved to allow for more accurate detection of noise at even higher SNR values, but at the expense of falsely classifying clean data as noisy. In moving the threshold in order to more accurately detect suspect signals, more clean signals would be at risk of being discarded. Where to set the boundary is therefore dependent on the nature of data collection. If a large amount of redundant data is available, then discarding clean data may not be problematic. Otherwise, an increase in discarded signals will augment the rate of re-collection, but will increase the overall sensitivity of the data.

## **Chapter 6. Comparison to SVM**

### **6.1 Purpose**

The purpose of this investigation was to compare the effectiveness of PANDA compared to another Any-Source Solution developed through the CleanEMG project – a one-class support vector machine (SVM) [18].

A support vector machine is a learning model that recognizes patterns in datasets using associated algorithms. Typically, two or more sets of data are used to train the SVM to sort test data into one of the datasets. The datasets are not easily separated with a linear function, therefore they are mapped to a higher dimensional space using a kernel function, where they can be separated by a hyperplane.

In a one-class SVM, the machine determines whether or not the test data belongs to the training data set. Only one set of data is used to train the SVM. For quality assessment of sEMG, a clean data set is used. The training data is input in the form of feature vectors and these vectors are mapped to a higher-dimensional space. The aim of the SVM is then to find a hyperplane that best separates the training data from the origin in the space. Mapping this hyperplane back into original space yields a tight volume of space enclosing the training data. Optimization methods are used to find the best solution for this. Once trained, the volume surface can be used as an outer bound for the class defined by the training data. Only test data that falls inside this volume is considered to be part of that class. Thus, in this work, any test data that falls outside the volume would be classed as noisy.

## **6.2 Method**

### **6.2-1 Recorded data collection**

The data collected in Sections 4.1-1 and 5.2-1 was used to complete the comparison of the SVM to PANDA. The baseline/training data consisted of data collected as described in Chapter 5 from the 12 participants ( $N_{BL} = 587$ ) and the test data consisted of the data collected as described in Chapter 4 from the 12 participants ( $N_{TE} = 290$ ).

### **6.2-2 Record contamination**

Record contamination followed the same process as outlined in Chapter 4, using the five contamination types (three separate and two blended). Copies of the clean test data set were created for each noise type and artificially contaminated to achieve a range of SNR values. The SNR values were chosen to focus on behaviour around the transition points for each noise type, as identified in Chapter 5: -10 to 10dB for PL and MA; 15 to 34dB for SAT; 0 to 15dB for the blended contamination types.

### **6.2-3 Testing method**

PANDA was configured as it was in Chapter 5 with i) a full feature set, ii)  $L = 5$  bins, and iii) a boundary setting of three standard deviations from the mean. All 587 data records were used to establish the boundary.

The SVM was configured as outlined by Fraser et al. [18]. In fact, the original system developed by Fraser was used in this work. A soft margin classifier was implemented with a Gaussian radial basis kernel and the method of Lagrange multipliers was used in optimization. Also as suggested by Fraser et al, the following six features were used as inputs to the SVM: 1) 10-bin amplitude histogram (first bin); 2) 10-bin amplitude

histogram (last bin); 3) Mean absolute value; 4) Willison amplitude; 5) 10-bin power spectrum (first bin); and 6) 10-bin power spectrum (second bin).

Note that the SVM and PANDA both use mean absolute value and the ten-bin histogram (or normality) feature, but differ with respect to the remainder of their feature sets. Also, PANDA does not separate the 10-bin histogram feature into the first and last bin; it combines the values for both bins into one feature value.

Once PANDA and the SVM were configured, the baseline/training data was processed through each. For PANDA, noise factors for each of the baseline signals were generated and the mean and boundary settings were established. For the SVM, the baseline data was implemented as training data. Once set up, the first run of test data was conducted for each tool; this used the test data that was known to be clean. The succeeding runs used the noisy data sets.

### 6.3 Results

In testing the clean data set, PANDA and the SVM were both 97% accurate in determining the data was clean; thus the probability of a false positive or false alarm for both was 3.0%. Figure 24 illustrates the results of testing the clean test data set with PANDA. The mean of the baseline data set ( $\mu \approx 56$ ) is depicted by the black line and the mean plus three standard deviations ( $\mu + 3\sigma = 108$ ) boundary line is shown in red. The blue and green data points represent the  $F_N$  values of the baseline data set and clean test data set, respectively; those points falling below the boundary line are considered clean. Here, 280/290, or 97%, of the points fell below the boundary. Figure 25 illustrates the results of testing the clean test data set with the SVM. The resulting data points were

scaled to  $[-1 \ 1]$ ; any points falling above zero are considered clean, while those falling below zero are considered contaminated. Here, 282/290, or 97%, of the points fell above zero.

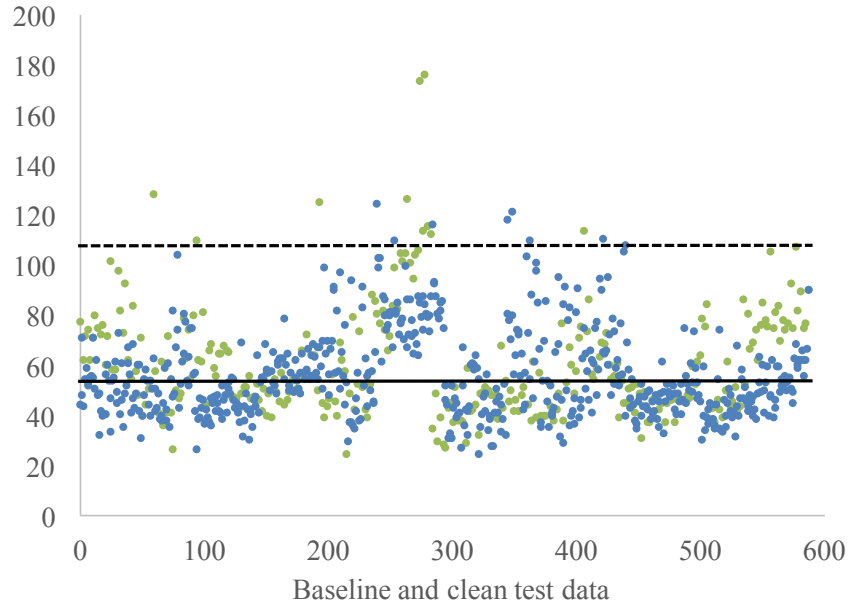


**Figure 24: Results of testing the baseline data set (blue points) and clean data set (green points) with PANDA**

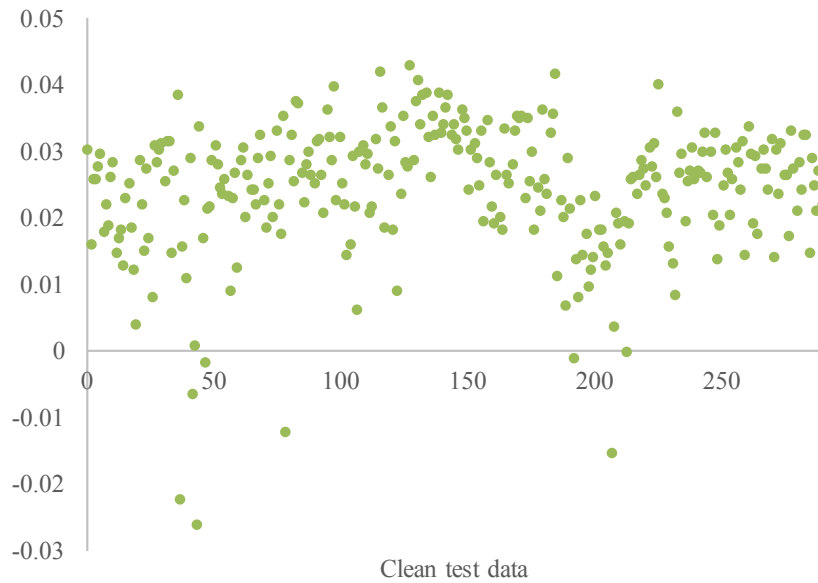


**Figure 25: Results of testing clean test data set in the SVM**

The following figures depict the sensitivity for both PANDA and the SVM for the five noise types (separate and blended) at varying SNR amounts. The figures display each tool's % sensitivity in detecting contaminated records as a function of SNR. The SNR values were chosen to illustrate the behavior of PANDA and SVM around the transition points found in the Chapter 5 investigation; therefore, each figure shows the transition points of each tool.

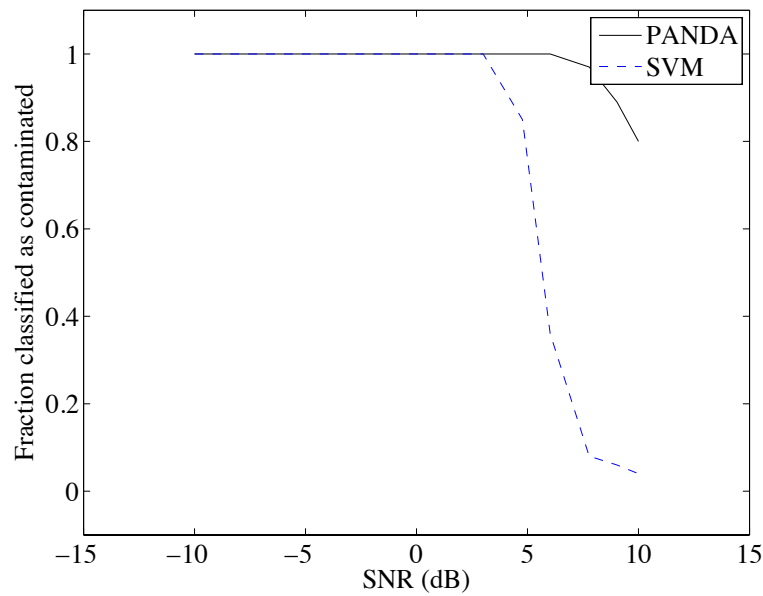


**Figure 26: Comparison of any-source solutions testing power line interference (recall  $P(\text{FA}) = 3.0\%$ ).**

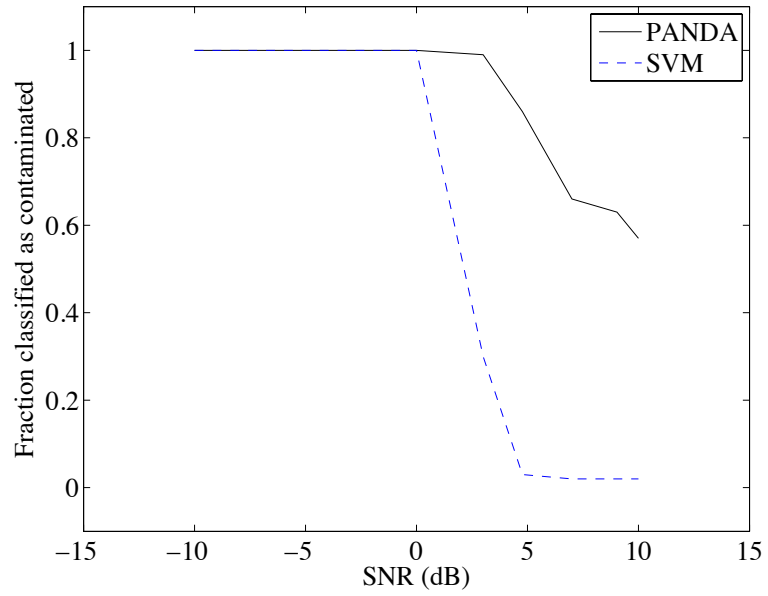


**Figure 27: Comparison of any-source solutions testing motion artifact (recall  $P(\text{FA}) = 3.0\%$ ).**

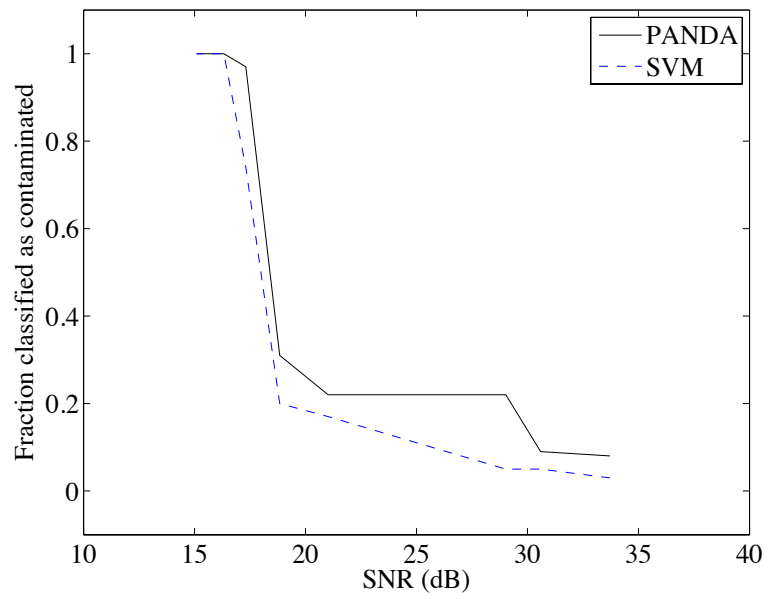


**Figure 28: Comparison of any-source solution testing saturation (recall  $P(\text{FA}) = 3.0\%$ ).**

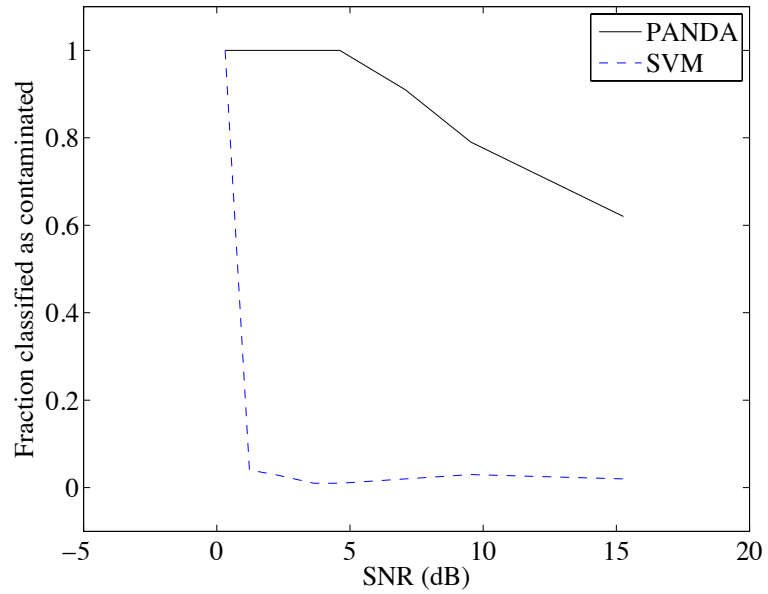


**Figure 29: Comparison of any-source solutions testing the combination of power line interference and motion artifact (recall  $P(\text{FA}) = 3.0\%$ ).**

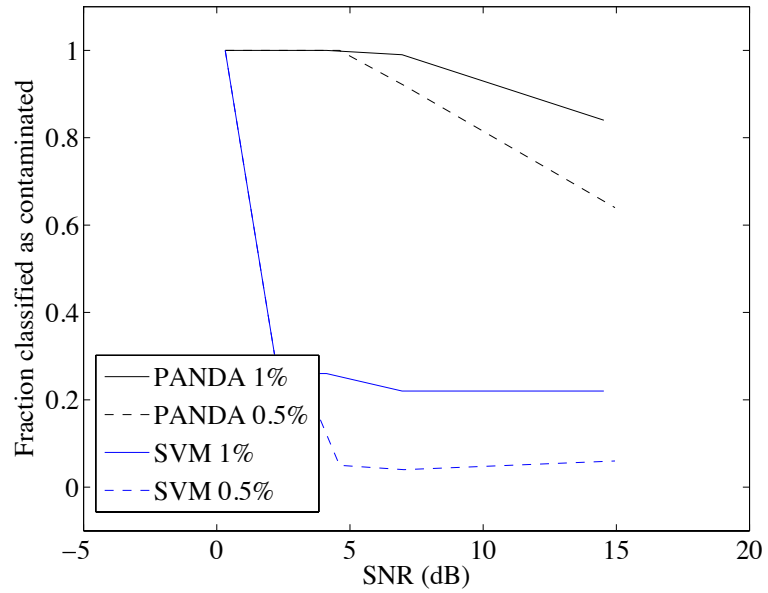


**Figure 30: Comparison of any-source solution testing the combination of power line interference, motion artifact, and saturation (recall  $P(\text{FA}) = 3.0\%$ ).**

## 6.4 Discussion

The effectiveness of PANDA was compared to another quality assessment tool – a one-class support vector machine (SVM) [18]. The same baseline and test data sets, comprised of recorded data, were used to compare the two tools. The same configuration used in Chapter 5 for PANDA was implemented here and the SVM was configured as detailed in the work completed by Fraser et al. [18]. Five noise types (three separate and two blended) were investigated.

The probability of a false alarm was found to be 3.0% for both PANDA and the SVM, which means that each method was 97% accurate in identifying clean signals. Both tools also performed well in identifying noisy data for all noise types, but the transition point about which performance started to degrade occurred at higher SNR values for PANDA in all cases. In this regard, PANDA outperformed the SVM since it was capable of detecting noisy data for signals with noise levels too small to be accurately detected by the SVM.

As depicted in Figure 26 for PL, both tools performed with 100% sensitivity until approximately 3dB. The SVM then experienced a steep transition point, while PANDA was able to maintain its perfect sensitivity until almost 7dB, where it then declined more slowly than the SVM.

As depicted in Figure 27 for MA, both tools performed with 100% sensitivity until approximately 0dB. The SVM then experienced a steep transition point, while PANDA was able to maintain near perfect sensitivity until 3dB, where it then declined more slowly than the SVM.

As depicted in Figure 28 for saturation, both tools performed with 100% sensitivity until approximately 17dB. Both tools then experienced steep transition points, although PANDA maintained slightly higher sensitivity, than the SVM, during the decline and plateau at 20dB.

As depicted in Figure 29 and Figure 30 for both blended types of noise, both tools performed with 100% sensitivity until 0dB. The SVM then experienced steep transition points for both blended types of noise, while PANDA was able to maintain near perfect sensitivity until 5dB and 7dB, respectively. In the second blend (where the saturation was added), the SVM and PANDA achieved better results for the records where 1% of saturation was added, compared to the 0.5% addition. It is also important to note that both PANDA and the SVM performed better with the second type of blended noise that included saturation. Both tools proved to be highly sensitive to instrumentation saturation.

While the results of this investigation indicate that PANDA outperforms the SVM, two considerations are worth noting. First, optimal configurations for each method were used in this comparison, but optimization was determined via different studies. PANDA's configuration was tuned through the work as detailed in the previous chapters and the SVM's configuration was tuned in the work completed by Fraser et al. [18]. Different features were considered in the optimizing processes so different feature sets emerged as optimal. This may account for PANDA's higher sensitivity. Future work should investigate the SVM's performance using the same feature set as PANDA and vice versa. It is worth noting that while there is utility in comparing each method using the same feature set, comparing the methods with feature sets specifically chosen to

optimize the performance of each may be more practical since this reflects how each would be actually used. It is also worth noting that PANDA implemented a subset of the features that were tested for optimization of the SVM.

Second, even though PANDA performed very well in this investigation, it may be worth implementing an SVM-like strategy with PANDA to further improve its performance. Unlike the SVM, PANDA was not designed to be a pattern classifier. PANDA produces a single metric for each of the test inputs; this work made PANDA into a pattern classifier using a threshold value for the metric as a classification boundary between clean and noisy records. Nevertheless, the algorithmic steps in PANDA do provide multidimensional error factors – one for each attribute. It may be possible to perform a multi-dimensional space calculation similar to the SVM to improve on PANDA's performance. Regardless, the comparison presented here supports PANDA's utility in quality assessment of sEMG.

## Chapter 7. Conclusion

The main purpose of this work was to modify an existing quality assessment algorithm to work with surface electromyography signals. PANDA was originally used for data mining software metrics, and the aim of this work was to modify it for incorporation into the CleanEMG project [2]. To accomplish this, the work was divided into three developmental parts: 1) configuration 2) in-use with simulated baseline data, 3) in-use with recorded data, and a concluding validation component comparing PANDA with SVM.

The first part used simulated sEMG data as the baseline and test data to verify the functionality of the algorithm and establish a suitable configuration. The configuration factors included i) feature set, ii) number of baseline signals, iii) number of bins, and iv) boundary setting. This yielded poor results as it became evident that the simulated data was sufficiently dissimilar to the recorded data to perform proper comparisons. Understanding the limitations of the simulated data at the time of testing led to a continuation of the work using recorded data as both the baseline and test data sets. The configuration of the algorithm for this portion of the work followed some of the original findings except a) a ninth feature, normality, was added to the feature set and b) receiver operating characteristic testing was performed to identify a suitable i) bin number and ii) boundary setting.

PANDA was evaluated with five noise types, separate and blended combinations of power line noise, motion artifact and saturation, at varying levels.

Results of the testing first indicated that PANDA was able to accurately identify clean signals with P(FA) of 9.1% or 4.2%, without the inclusion of an outlier. Next, it was

demonstrated that PANDA was able to correctly identify all types and combinations of noise at varying contamination levels with 100% sensitivity until a transition point: 7dB for PL interference; 3dB for MA; 17dB for saturation; 7dB for the combination of PL interference and MA; and 9dB for the combination of all three noise types.

A comparison was performed between PANDA and another similar quality-assessment tool – the Support Vector Machine. The study demonstrated that PANDA performed better in the identification of all noise types compared to the SVM. PANDA moderately out-performed the SVM in identifying the separate noise types (PL, MA, and SAT), but markedly out-performed the SVM in the identification of the blended noise types (PL-MA and PL-MA-SAT). The noise detection techniques that are part of the Clean EMG project [2] focus on single sources of noise, so PANDA's ability to accurately detect signals with multiple noise sources is a tremendous result.

## **7.1 Recommendations**

PANDA should be configured, to detect noisy data, as follows: i) a full feature set, ii)  $N \geq 600$  baseline records (600 records were used here, but preliminary work demonstrated that more baseline records worked well also), iii)  $L = 5$  bins with equal bin widths, and iv) boundary setting of three standard deviations from the mean. To execute PANDA, a set of baseline data should first be collected and its cleanliness should be verified through alternate quality assessment algorithms, such as those included in the CleanEMG project [2]. Then, PANDA should be configured and tuned and finally, the test data can be collected and run through PANDA.

PANDA was successful at differentiating between clean and contaminated sEMG signals, but further work could be completed to make the simulated baseline data alternative more viable. The simulated data set that was established for the preliminary testing could be given more attention. A thorough investigation of Myosim [25] can be completed in order to produce a simulated data set that better represents a recorded data set. One limitation, currently inherent to Myosim, is the lack of baseline noise in the simulated output. Work should be done to characterize baseline noise in sEMG and add it to the simulated output. Two other limitations, based on the experimentation design of this work, are: the potentially poor setup of the physiological parameters of Myosim and the homogeneity of the baseline data. If a simulated data set could be established that worked as well as the recorded data set, it would greatly reduce the time needed to use PANDA – both in terms of collection and computational time. An improved simulated data set could be realized through better setup of the Myosim parameters and through increased variability of the baseline data, in order to increase heterogeneity.

Future work could also be completed to better understand the functionality of PANDA and, perhaps, improve its performance. A visualization tool could be implemented with the PANDA simulator to better represent the binning process of the algorithm. As noted in Chapter 3, there are different binning methods available to the algorithm. A visualization tool may provide the user with more insight into the best binning method. The visualization tool might also provide insight into how noise affects features, which might suggest additional features for consideration. Also, an investigation should also be done to determine if PANDA's ability to detect combinations of contamination could be taken one step further to dissect the contamination into its different noise sources.

PANDA was shown to work very well in detecting blended types of contamination, so it may be valuable to the researcher to understand the make-up of the blend of the contamination, especially if the researcher aims to remove the noise.

Future work could also be done to examine possible applications outside noise detection for the algorithm. For instance, PANDA may be an ideal tool for neuromuscular disease progression analyses. Baseline data from a patient could be established and test data would be comprised of data from a future visit. If the test data was identified as “suspect”, the clinician would be notified that the patient’s disease was progressing, either negatively or positively. While the current work proposes a feature set well suited to identify noisy data, further work could investigate the possibility of using alternate feature sets for other applications.

Currently, PANDA is used in an offline, post-processing capacity. Researchers are able to use PANDA as a quality assessment tool after their data is collected. This could remain a viable use for PANDA with improved operation being realized by the continued addition of clean baseline data; these additions should bolster PANDA’s sensitivity. PANDA could also be incorporated in an embedded system or used, in a software application, while researchers are collecting their data. If PANDA was incorporated during the data collection process, researchers could automatically know to throw away noisy records and collect more while their apparatus was in place. PANDA has the potential for great utility for quality assessment of sEMG.

## References

- [1] J. Van Hulse, T.M. Khoshgoftaar, H. Huang, "The pairwise attribute noise detection algorithm," *Knowl Inf Syst*, vol. 11, no. 2, pp. 171-190, 2007.
- [2] A.D. Chan, D. MacIsaac, "CleanEMG: Assessing the quality of EMG signals", *34<sup>th</sup> Conference of the Canadian Medical & Biological Engineering Society and Festival of International Conferences on Caregiving, Disability, Aging and Technology*, Toronto, Canada, 69826, pp. 1-4, 2011.
- [3] R. Merletti and P. Parker, *Electromyography - Physiology, Engineering, and Noninvasive Applications*, John Wiley & Sons, 2004.
- [4] J. E. Bolek, "Use of Multiple-Site Performance-Contingent SEMG Reward Programming in Pediatric Rehabilitation: A Retrospective Review." *Applied Psychophysiology and Biofeedback*, vol. 21, no. 3, pp. 263-272, September 2006.
- [5] "Albert's graduate students conduct research study with Fredericton police force." *Faculty of Kinesiology Newsletter* (April, 2014), Issue 1, pp. 1.
- [6] "Pocket Ergometer for EMG (Electromyography)." Internet: <http://www.thehumansolution.com/emg.html>. [02-04-2016].
- [7] A. Pause, "MYO changes human-technology interaction." Internet: Available <http://engineerthefuture.ca/myo-changes-human-technology-interaction/> [04-02-2016]
- [8] C. Grönlund, K. Roeleveld, A. Holtermann, J. S. Karlsson, "On-line signal quality estimation of multichannel surface electromyograms," *Med. Biol. Eng. Comput.*, vol. 43, pp. 357-364, 2005.
- [9] E. A. Clancy, E. L. Morin, R. Merletti, "Sampling, noise-reduction and amplitude estimation issues in surface electromyography," *J. Electr. Kinesiol.*, vol. 12, pp. 1-16, 2002.
- [10] E. Huigen, A. Peper, C. A. Grimbergen, "Investigation into the origin of noise of surface electrodes", *Med. Biol. Eng. Comput.*, vol. 40, pp. 332-338, 2002.
- [11] R. Merletti, G. Rau, C. Disselhorst-Klug, D.F. Stegerman, and G. M. Hägg, "SENIAM." Internet: <http://www.seniam.org>. [03-03-2015]
- [12] N. Abser, D. MacIsaac, G. Fraser, A. D. C. Chan, J. R. Green, "CleanEMG: Quantifying power line interference in surface EMG signals", *34th Conference of the Canadian Medical & Biological Engineering Society and Festival of*

*International Conferences on Caregiving, Disability, Aging and Technology*, Toronto, Canada, 69825, pp. 1-4, 2011.

- [13] G. D. Fraser, A. D. C. Chan, J.R. Green, D. MacIsaac, "Detection of ADC clipping, quantization noise, and amplifier saturation in surface electromyography," in *Medical Measurements and Applications Proceedings (MeMeA)*, 2010 IEEE International Symposium on, 2012, pp. 162-166.
- [14] G. D. Fraser, A. D. C. Chan, J.R. Green, D. MacIsaac, " Removal of electrocardiogram artifacts in surface electromyography using a moving average method," in *Medical Measurements and Applications Proceedings (MeMeA)*, 2010 IEEE International Symposium on, 2012, pp. 128-131.
- [15] A. Arvidsson, A. Grassino, L. Lindström, "Automatic selection of uncontaminated electromyogram as applied to respiratory muscle fatigue," *J. Appl. Physiol.: Respirat. Environ. Exercise Physiol.* vol. 56(3), pp. 568-575, 1984.
- [16] D. T. Mewett, K. J. Reynolds, and H. Nazeran, "Reducing power line interference in digitized electromyogram recordings by spectrum interpolation," *Medical & Biological Engineering & Computing*, vol. 42, no. 4, pp. 524-531, 2004.
- [17] A. Bartolo, R. R. Dzwonczyk, C. Roberts, E. Goldman. "Description and validation of a technique for the removal of ECG contamination from diaphragmatic EMG signal," *Medical and Biological Engineering and Computing*, vol. 34, no. 1, pp. 76-81, 1996.
- [18] G.D. Fraser, A.D.C. Chan, J.R. Green, "Automated biosignal quality analysis for electromyography using a one-class support vector machine", *IEEE Transactions on Instrumentation and Measurement*, vol. 63., no. 12, pp. 2919-2930, 2014.
- [19] T.M. Khoshgoftaar, J. Van Hulse, "Empirical Case Studies in Attribute Noise Detection," *IEEE Trans on Systems, Man, and Cybernetics-Part C: Applications and reviews*, vol. 39, no. 4, pp. 379-388, July 2009.
- [20] De Luca, C. J., L. D. Gilmore, M. Kuznetsov, S. H. Roy, "Filtering the surface EMG signal: Movement artifact and baseline noise contamination", *Journal of Biomechanics*, vol. 43, pp. 1573-1579, 2010.
- [21] D. Gradolewski, P. M. Tojza, J. Jaworski, D. Ambroziak, G. Redlarski, M. Krawczuk, "Arm EMG Wavelet-Based Denoising System," *Mechatronics – Ideas for Industrial Applications*, vol. 317, pp 289-296, 2015.
- [22] S. N. Kale, S. V. Dudul, "Intelligent Noise Removal from EMG Signal Using Focused Time-Lagged Recurrent Neural Network," *Appl. Computational Intelligence and Soft Computing*, vol. 2009, pp. 1-12, 2009.

- [23] X. Zhang and P. Zhou, "Sample entropy analysis of surface EMG for improved muscle activity onset detection against spurious background spikes," *Journal of Electromyography and Kinesiology*, vol. 22, no. 6, pp. 901-907, Dec., 2012.
- [24] C. Sinderby, L. Lindström, A.E. Grassino, "Automatic assessment of electromyogram quality," *L. Appl. Physiol.*, vol. 79(5), pp. 1803-1815, 1995.
- [25] D. MacIsaac, D. Rogers, A. Bhandarkar, "Simulating Myoelectric Signals with a Finite-Length Model Muscle," Fredericton, 2007.
- [26] K. C. McGill, "A model of the muscle action potential for describing the leading edge, terminal wave, and slow afterwave", *IEEE Trans Biomed Eng.*, vol. 48(12), pp. 1357-65, Dec. 2001.
- [27] R. Dawson, "How significant is a boxplot outlier?", *Journal of Statistics Education*, vol. 19(2), pp. 1-12, 2011.

## Appendix

### Example of PANDA

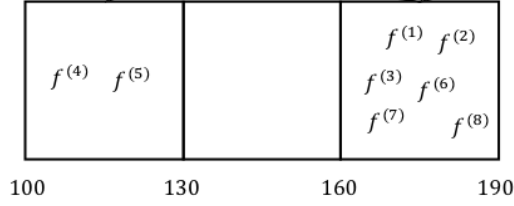
N = 8 records  
M = 3 features  
L = 3 bins

1

Establish 8 feature vectors

$$f^{(1)} = \begin{bmatrix} 190 \\ 27 \\ 100 \end{bmatrix} f^{(2)} = \begin{bmatrix} 170 \\ 28 \\ 61 \end{bmatrix} f^{(3)} = \begin{bmatrix} 161 \\ 38 \\ 68 \end{bmatrix} f^{(4)} = \begin{bmatrix} 100 \\ 26 \\ 55 \end{bmatrix} f^{(5)} = \begin{bmatrix} 105 \\ 25 \\ 40 \end{bmatrix} f^{(6)} = \begin{bmatrix} 173 \\ 50 \\ 58 \end{bmatrix} f^{(7)} = \begin{bmatrix} 173 \\ 39 \\ 89 \end{bmatrix} f^{(8)} = \begin{bmatrix} 180 \\ 55 \\ 68 \end{bmatrix}$$

Cluster set  $C_1$  with  
L = 3 bins.



$$\begin{aligned} \text{Bin width} &= \frac{\max_{obs} - \min_{obs}}{L} \\ \text{Bin width} &= \frac{190 - 100}{3} \\ \text{Bin width} &= 30 \end{aligned}$$

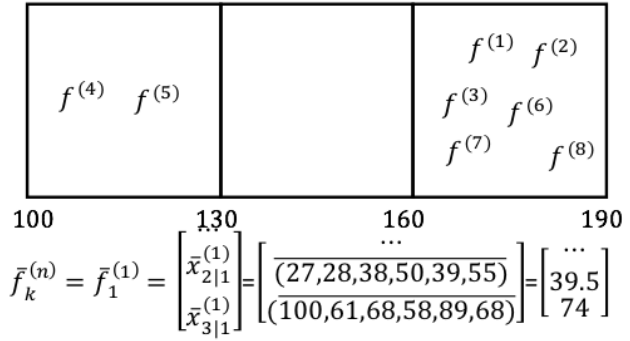
2

Bin 8 feature vectors using  
equal-width binning  
method for the first feature

Cluster set  $C_1$

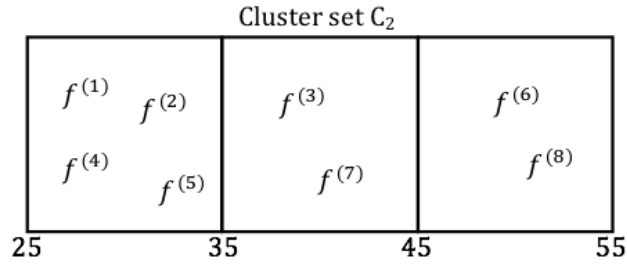
3a

Cluster set  $C_1$  shows  
the binning of the 8  
feature sets



$$\bar{f}_k^{(n)} = \bar{f}_1^{(1)} = \begin{bmatrix} \bar{x}_{2|1}^{(1)} \\ \bar{x}_{3|1}^{(1)} \end{bmatrix} = \begin{bmatrix} \dots \\ \overline{(27, 28, 38, 50, 39, 55)} \\ \overline{(100, 61, 68, 58, 89, 68)} \end{bmatrix} = \begin{bmatrix} \dots \\ 39.5 \\ 74 \end{bmatrix}$$

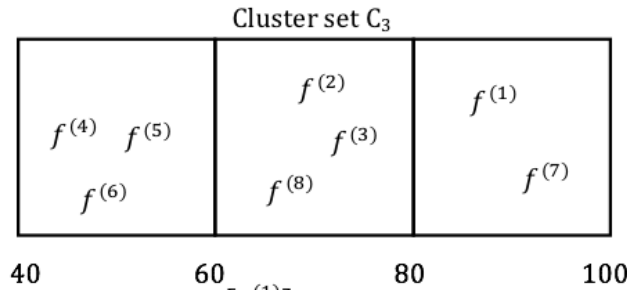
$$\bar{f}_k^{(n)} = \bar{f}_1^{(n)} = \begin{bmatrix} \bar{x}_{2|1}^{(1)} \\ \bar{x}_{3|1}^{(1)} \end{bmatrix} = \begin{bmatrix} \dots \\ \overline{(27, 28, 38, 50, 39, 55)} \\ \overline{(100, 61, 68, 58, 89, 68)} \end{bmatrix} = \begin{bmatrix} \dots \\ 11.3 \\ 16.72 \end{bmatrix}$$



$$\bar{f}_k^{(n)} = \bar{f}_2^{(1)} = \begin{bmatrix} \bar{x}_{1|2}^{(1)} \\ \dots \\ \bar{x}_{3|2}^{(1)} \end{bmatrix} = \begin{bmatrix} \overline{(190,170,100,105)} \\ \dots \\ \overline{(100,61,55,40)} \end{bmatrix} = \begin{bmatrix} 141.25 \\ \dots \\ 64 \end{bmatrix}$$

$$\bar{\bar{f}}_k^{(n)} = \bar{\bar{f}}_2^{(n)} = \begin{bmatrix} \bar{\bar{x}}_{1|2}^{(1)} \\ \dots \\ \bar{\bar{x}}_{3|2}^{(1)} \end{bmatrix} = \begin{bmatrix} \overline{\overline{(190,170,100,105)}} \\ \dots \\ \overline{\overline{(100,61,55,40)}} \end{bmatrix} = \begin{bmatrix} 45.53 \\ \dots \\ 25.57 \end{bmatrix}$$

|           |                                                                     |
|-----------|---------------------------------------------------------------------|
| <b>3b</b> | Cluster sets $C_2$ and $C_3$ show the binning of the 8 feature sets |
|-----------|---------------------------------------------------------------------|



$$\bar{f}_k^{(n)} = \bar{f}_3^{(1)} = \begin{bmatrix} \bar{x}_{1|3}^{(1)} \\ \bar{x}_{2|3}^{(1)} \\ \dots \end{bmatrix} = \begin{bmatrix} \overline{(190,173)} \\ \overline{(27,39)} \\ \dots \end{bmatrix} = \begin{bmatrix} 181.25 \\ 33 \\ \dots \end{bmatrix}$$

$$\bar{\bar{f}}_k^{(n)} = \bar{\bar{f}}_3^{(n)} = \begin{bmatrix} \bar{\bar{x}}_{1|3}^{(1)} \\ \bar{\bar{x}}_{2|3}^{(1)} \\ \dots \end{bmatrix} = \begin{bmatrix} \overline{\overline{(190,173)}} \\ \overline{\overline{(27,39)}} \\ \dots \end{bmatrix} = \begin{bmatrix} 12.02 \\ 8.49 \\ \dots \end{bmatrix}$$

|   |                                                                                                                                                    |
|---|----------------------------------------------------------------------------------------------------------------------------------------------------|
| 4 | Calculate the expected value vector ( $\bar{f}_k^{(n)}$ ) and standard deviation vector ( $\bar{\sigma}_k^{(n)}$ ) for the 1 <sup>st</sup> feature |
|---|----------------------------------------------------------------------------------------------------------------------------------------------------|

$$\Delta^{(n)} = \frac{|\bar{f}^{(n)} - f^{(n)}|}{\bar{\sigma}^{(n)}}$$

$$= \begin{bmatrix} \dots & 141.25 & 181.25 \\ 39.5 & \dots & 33 \\ 74 & 64 & \dots \end{bmatrix} - \begin{bmatrix} 190 \\ 27 \\ 100 \end{bmatrix} / \begin{bmatrix} \dots & 45.53 & 12.02 \\ 11.3 & \dots & 8.49 \\ 16.72 & 25.57 & \dots \end{bmatrix}$$

|   |                                                                                          |
|---|------------------------------------------------------------------------------------------|
| 5 | Calculate the deviation matrix ( $\Delta^{(n)}$ ) for the 1 <sup>st</sup> feature vector |
|---|------------------------------------------------------------------------------------------|

$$\Delta^{(n)} = \begin{bmatrix} \dots & 1.07 & 0.71 \\ 1.1 & \dots & 0.71 \\ 1.56 & 1.41 & \dots \end{bmatrix}$$

|   |                                                                      |
|---|----------------------------------------------------------------------|
| 6 | Calculate the total deviation for the 1 <sup>st</sup> feature vector |
|---|----------------------------------------------------------------------|

$$E_m^{(1)} = \begin{bmatrix} 1.78 \\ 1.81 \\ 2.97 \end{bmatrix}$$

|   |                                                                    |
|---|--------------------------------------------------------------------|
| 7 | Start of the calculation for the sum of deviations across features |
|---|--------------------------------------------------------------------|

$$F_N = \begin{bmatrix} 1.78 \\ 1.81 \\ 2.97 \end{bmatrix} + E_m^{(2)} + \dots + E_m^{(8)}$$

## **Curriculum Vitae**

Gillian Lesley Phillips

University of New Brunswick

2012 – present

Masters in Electrical and Computer Engineering

2007-2012

Bachelors of Electrical Engineering

Biomedical Engineering Option

Minor in Mathematics

Diploma of Technology Management and Entrepreneurship

Conference Presentations:

The XX ISEK Conference – Rome, Italy 15<sup>th</sup> – 18<sup>th</sup> July 2014

Abstract title: Pairwise Attribute Noise Detection Applied to Surface EMG