# Radiometric Identification of Long Term Evolution Transmitters

by

**Frédéric Demers**, **B.Eng.**

A thesis submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

**Master of Applied Science in Electrical and Computer Engineering**

Ottawa-Carleton Institute for Electrical and Computer Engineering (OCIECE)
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada, K1S 5B6
April, 2013

The undersigned hereby recommends to the
Faculty of Graduate and Postdoctoral Affairs
acceptance of the thesis

# Radiometric Identification of Long Term Evolution Transmitters

submitted by **Frédéric Demers**, **B.Eng.**

in partial fulfillment of the requirements for the degree of

**Master of Applied Science in Electrical and Computer Engineering**

---

Professor Howard Schwartz, Chair,
Department of Systems and Computer Engineering

---

Professor Marc St-Hilaire, Thesis Supervisor

Carleton University
April, 2013

# Abstract

This thesis investigates whether the radiometric identification of Long Term Evolution (LTE) transmitters is possible using commercial off-the-shelf hardware and support vector machine (SVM) classifiers. The identification is based on unique modulation characteristics exhibited by the transmitters, originating from minute imperfections introduced during radio hardware manufacturing. In these experiments, the Agilent vector signal analysis (VSA) software and the Agilent PXA spectrum analyzer are used to extract radiometric properties from several LTE base stations, known as evolved Node B (eNB). The open source SVM library *libsvm* performs the classification using 13 emitter-specific coefficients extracted by the VSA software.

It is shown that when SVM parameters are optimized using grid search, and the training bin contains no less than 45 vectors, re-identification success exceeds 98% for two Ottawa-based providers. Furthermore, the use of feature scoring and weighting is investigated and shown to result in faster convergence for small training bins and slightly better re-identification accuracy.

# Acknowledgments

I would first like to express my sincere gratitude to my supervisor Professor Marc St-Hilaire. Your accurate and rapid feedback, unfailing encouragement and personal guidance have been a source of inspiration throughout this long effort. I wish to also thank the professors of the Ottawa-Carleton Institute for Electrical and Computer Engineering (OCIECE) and School of Computer Science at Carleton University, who have shown remarkable knowledge and wisdom.

Furthermore, I would like to take this opportunity to thank the dedicated members of the Modern Communications Electronic Warfare (MCEW) team from Defence Research and Development Canada (DRDC) Ottawa, for supporting my research efforts in many ways. In particular, I wish to recognize Bruce Liao and Susan Watson: I was fortunate to be able to draw on your technical knowledge and commendable experience; your insightful advice has been precious throughout.

I also wish to recognize my many professional colleagues who exemplify some of the best qualities the Canadian workforce has to offer. I always welcome and appreciate your constructive feedback and support. I extend a special thanks to the managers and directors who afforded me time at critical points in my Master's coursework and during the writing of this thesis.

Lastly, and most importantly, I wish to dedicate this thesis to my family. To Sonia, you have spent so many days alone taking care of our daughter Annabel whilst I took time away to research, experiment and write. I cannot overstate that this project could not have been completed without your love and support.

To Annabel, you are a very lively and pleasant child. It has been difficult at times to prioritize my school work instead of playing with you during your first few years of life. Let the completion of this thesis serve as proof that if you work hard, you too can realize your dreams, and remember that I will be there to help you along the way.

# Table of Contents

# List of Tables

# List of Figures

# Nomenclature

**1G** first generation

**2G** second generation

**3G** third generation

**3GPP** Third Generation Partnership Project

**3GPP2** Third Generation Partnership Project 2

**4G** fourth generation

**API** application programming interface

**BCH** broadcast channel

**BPSK** binary phase shift keying

**BSCD** Bayesian step change detector

**CDMA** code division multiple access

**CTE** common tracking error

**CSV** comma-separated values

**DAC** digital-to-analog converter

**DAG** directed acyclic graph

**dB** decibel

**DC** direct current

**DRDC** Defence Research and Development Canada

**DT** decision tree

**eNB** evolved Node B

**E-UTRA** evolved UMTS terrestrial radio access

**EVM** error vector magnitude

**FDD** frequency division duplexing

**FFT** fast Fourier transform

**GLRT** generalized likelihood ratio test

**GSM** Global System for Mobile communication

**HSPA** high speed packet access

**HTML** HyperText Markup Language

**IEEE** Institute of Electrical and Electronics Engineers

**IMEI** international mobile equipment identity

**IP** Internet protocol

**IQ** in-phase and quadrature

**KLT** Karhunen-Loève transformation

**kNN** $k$-nearest neighbours

**LAN** local area network

**LDA** linear discriminant analysis

**LTE** Long Term Evolution

**MAC** media access control

**MCEW** Modern Communications Electronic Warfare

**MDL** minimum description length

**MIMO** multiple inputs multiple outputs

**mRMR** minimum-redundancy-maximum-relevance

**OCIECE** Ottawa-Carleton Institute for Electrical and Computer Engineering

**OFDM** orthogonal frequency division multiplexing

**OFDMA** orthogonal frequency division multiple access

**OSTP** orthogonal frequency division multiplexing (OFDM) symbol transmit power

**PARADIS** Passive RAdiometric Device Identification System

**PCA** principal component analysis

**PCI** physical cell identity

**PNN** probabilistic neural network

**PBCH** physical broadcast channel

**PDCCH** physical downlink control channel

**PDSCH** physical downlink shared channel

**ppm** parts-per-million

**PRI** pulse repetition interval

**PSS** primary synchronization signal

**QAM** quadrature amplitude modulation

**QPSK** quadrature phase shift keying

**RB** resource block

**RBF** radial basis function

**RF** radio frequency

**RMS** root mean square

**RPM** revolutions per minute

**RS** reference signal

**RSEVM** received signal error vector magnitude

**RSS** received signal strength

**RSSI** received signal strength indication

**RSRP** reference signal received power

**RSRQ** reference signal received quality

**RSTP** reference signal transmit power

**SC-FDM** single-carrier frequency division multiplexing

**SC-FDMA** single-carrier frequency division multiple access

**SEI** specific emitter identification

**SIM** subscriber identity module

**SMS** short messaging service

**SNR** signal-to-noise ratio

**SQL** structured query language

**SSS** secondary synchronization signal

**SVM** support vector machine

**TDD** time division duplexing

**TDMA** time division multiple access

**UE** user equipment

**UMB** Ultra-Mobile Broadband

**UMTS** Universal Mobile Telecommunications System

**USIM** universal subscriber identity module

**USRP** universal software radio peripheral

**VBA** Visual Basic for Applications

**VSA** vector signal analysis

**VPS** virtual private server

**WCDMA** wideband code division multiple access (CDMA)

**WEKA** Waikato Environment for Knowledge Analysis

**WLA** wireless link analysis

# Chapter 1

# Introduction

## 1.1 Motivation and Applications

The ability to uniquely identify communication emitters is essential to many defence and security applications. Wireless service providers have a keen interest in ensuring that only legitimate users are granted network access. This is accomplished in a variety of ways, some as simple as user authentication, or hard-coded identification strings provided by the user equipment (UE), which are queried against an equipment database. Since users are granted services and privileges from untethered locations, authentication is more difficult [1]. Network operators have little control over the UE when verifying its authenticity. The device's hardware identification string is often not sufficiently reliable because it can easily be spoofed or cloned [2–4].

Conversely, users may wish to verify the access point or base station to which they connect legitimately belongs to the service provider, rather than a rogue device impersonating the providers' cellular equipment. The risks of impersonation in wireless network was discussed by Barbeau et al. in [5]. A typical attack involving a rogue access point is further described by Hall et al. in [6], and can result in the users' traffic being intercepted by the attacker.

Many modern cellular standards have struggled with cloned devices and faked equipment identities — e.g. international mobile equipment identity (IMEI) spoofing, in Global System for Mobile communication (GSM). However, due to imperfect manufacturing processes of radio frequency (RF) components, minute differences are introduced by emitters during modulation, even for emitters of the same make and model. It is therefore possible to uniquely distinguish emitters by characterizing

aspects of the transmitted signal, a technique fittingly named radiometric identification. Radiometric identification has been demonstrated as an effective way to defeat media access control (MAC) address cloning in 802.11 networks, or subscriber identity module (SIM) card cloning in second generation (2G) cellular networks [7].

Subscriber identity in third generation (3G) or fourth generation (4G) networks is cryptographically protected and much harder to clone. In fact, the user-to-universal subscriber identity module (USIM) link is protected by a shared secret stored securely in the USIM or provided interactively by the user. The USIM-to-terminal link is also protected by a shared secret [8]. Network operators can also detect when two users with the same USIM parameters access the network simultaneously, ensuring this type of attack is largely avoided. In 3G and 4G systems, the network provider equipment mutually authenticates with the user prior to the communication taking place [9]. In spite of these improvements, there are a number of practical applications which warrant the study of radiometric identification for the newer cellular protocols.

First, radiometric identification, also termed radio fingerprinting or specific emitter identification (SEI), can be used to conduct device and user tracking without the necessity to decrypt user identification strings. The need to discover a user's resource block allocation (logical-physical channel and time assignments) is also removed. As such, emitter tracking becomes possible for by-standers not privy to the communication messaging nor encrypted identifying strings, solely recognizing the unique modulation characteristics of the user's equipment. Conversely, radiometric identification of infrastructure emitters could help users make the determination if the base station is legitimate or is being impersonated by an attacker who may have succeeded in defeating the mutual authentication. Another attack, described by Meyer and Wetzel in [9], facilitates a man-in-the-middle attack to 3G handset by purposefully downgrading the network equipment to GSM in order to bypass mutual authentication. This type of attack could be prevented using properly implemented radiometric identification. Chapter 2 presents other applications for radiometric identification in radar systems and communication systems other than Long Term Evolution (LTE).

## 1.2 Research Objectives

In [2], Brik et al. demonstrated a very low error rate classifying IEEE 802.11/WiFi network interface cards, using characteristics of the modulated signal and a support

vector machine (SVM) classifier. Classification was conducted with five parameters collected by the Agilent vector signal analysis (VSA) software during a learning period: centre frequency offset, modulation constellation centre offset, average symbol error vector magnitude (EVM) and phase error, and SYNC correlation. Once the learning period was completed, the classifier algorithm attempted to match the parameters collected from an unknown emitter against the classes of emitters studied during the learning period. Brik et al. were able to uniquely identify transmitters out of 130 WiFi cards from the same manufacturer and using the same chip set, in excess of 99% accuracy. In this work, we wish to determine whether the modulation characteristics of LTE transmitters can be characterized with sufficient uniqueness as to re-establish their identity once a set of signature vectors has been gathered. The term radiometric identification is used henceforth to describe this activity.

### 1.2.1 Primary Research Objectives

The main objective of this research is to examine the suitability of the method proposed by Brik et al. against newer cellular communication systems and assess whether the re-identification success rates can be improved using a weighted classifier algorithm favouring the parameters with the most variance and least redundancy. More precisely, we will show that unique identification of LTE emitters is possible using the following steps:

- Record RF signals from different LTE transmitters with an Agilent PXA spectrum analyzer.
- Extract modulation characteristics from the recordings, for each transmitter, using the Agilent VSA software.
- Train the machine learning classifier with a subset of the parameter vectors to produce a class model.
- Predict the identity of emitters when presented an unknown parameter vector, using the class model.

### 1.2.2 Secondary Research Objectives

- Determine which parameters are most effective for accurate re-identification, using two feature scoring algorithms.

- Examine the effect of parameter weighting during emitter classification and re-identification.
- Study the impact of adverse signal quality on emitter identification.
- Compare the parameters chosen by Brik et al. against those most highly ranked by the feature scoring algorithms, as a special case of feature weighting.

## 1.3 Hypotheses

The hypothesis is made that it is possible to collect a set of modulation characteristics from an LTE base station's downlink signal, using advanced test instrumentation such as a spectrum analyser with VSA capability. Once the characteristics have been harvested, it is further hypothesized that a classifier algorithm, such as SVM, can accurately predict the identity of an LTE emitter when presented with an unknown feature vector. We further hypothesise that efforts towards assigning a heavier weight to the most important features will improve re-identification accuracy.

## 1.4 Contributions

This research activity aims to present novel contributions in the following areas:

- First successful radiometric identification of LTE transmitters using modulation characteristics collected by the VSA software. Many North American network operators have first launched LTE service during the Summer and Fall of 2011. As such, it has been challenging until now to conduct empirical research using emitters from active 4G networks. To the best of our knowledge, this is the first attempt at verifying this hypothesis. It is also hoped that the work completed here will be of use for many years to come.
- First successful use of a weighted classifier algorithm such as weighted-SVM to conduct the identification. Previous radiometric identification algorithms considered each feature evenly. While the success rates for IEEE 802.11/WiFi as reported by Brik et al. were remarkable, it is suspected that the use of a classifier algorithm that favours certain parameters will further enhance re-identification success rates.

- First thorough examination of the impact of adaptive modulation on re-identification. Unlike the work presented by Brik et al. in [2], LTE transmitters have the ability to adapt their modulation between quadrature phase shift keying (QPSK), 16 quadrature amplitude modulation (QAM) and 64 QAM in line with the quality of the radio channel between the UE and the base station, called evolved Node B (eNB) in LTE. It may be necessary to gather radiometric characteristics for each modulation schemes for re-identification to succeed.

## 1.5 Outline

The rest of this thesis is organized as follows. Chapter 2 provides background information useful to the reader concerning cellular standard evolution, LTE and SVMs. The last part of Chapter 2 provides a review of the research literature related radiometric identification and formal feature selection. Chapter 3 provides the theoretical notions applied in this research effort, presents the 13 coefficients extracted from the VSA, and details the steps of the analysis process applied to the data collected. Chapter 3 also presents the five weighting algorithms evaluated in this study. The empirical results and their significance are presented next, in Chapter 4. Finally, conclusions and recommendations for future work are offered in Chapter 5.

# Chapter 2

# Background and Literature Review

This chapter presents useful background information followed by an overview of recent findings published in research literature. Aspects of LTE protocol required to understand this thesis are discussed next. The necessity to conduct emitter discrimination and its most common approaches follow, along with amplifying details concerning the source of the uniqueness present in radio signals. A discussion on the reliability of radiometric identification in real-world applications appears next. Finally, an introduction to SVMs and a brief overview of formal feature selection concludes the chapter.

## 2.1 Long Term Evolution (LTE)

In this section, we provide background information that is useful for the reader throughout this thesis. We first start by looking at the cellular network evolution from first generation to today's fourth generation. Then, we provide the reader with an introduction to LTE specifications, modulation and multiplexing schemes relevant to this study.

### 2.1.1 Evolution towards Long Term Evolution

Research work for the $2^{nd}$ generation of mobile communication systems started in Europe in the early 1980s, and the complete system was ready for market in 1990. The most successful 2G system, called GSM is based upon time division multiple access (TDMA) in which eight users share a single narrowband radio channel. In North-America, service providers chose for the most part the competing code division

multiple access (CDMA) standard. These 2G systems replaced the 1st generation analogue cellular systems.

Due to the limited throughput offered by 2G systems, several research efforts were made in order to develop a third generation of cellular networks. Universal Mobile Telecommunications System (UMTS) was the predominant 3G standard globally and started commercial implementation around 2002 [10]. North-American network operators mostly opted for CDMA-2000 services but many converted at some point to the UMTS or its high speed packet access (HSPA)-based extensions.

Compared with earlier GSM networks, these UMTS systems provide much higher data throughput, typically in the range of 64 to 384 kbit/s, while the peak data rate for low mobility or indoor applications approaches 2 Mbit/s. With the improvements offered by HSPA, data rates of up to 7.2 Mbit/s are available in the downlink [10].

When UMTS was designed, the air interface was specified with a carrier bandwidth of 5 MHz. Wideband CDMA (WCDMA), the air interface chosen at that time, performed very well within this limit. Unfortunately, WCDMA does not scale well. If the bandwidth of the carrier is increased to sustain higher transmission speeds, the time between two transmission steps, or symbols, has to decrease. This results in transmissions being more vulnerable to multipath effects.

Instead of spreading one signal over the complete carrier bandwidth (e.g. 5 MHz), LTE transmits the data over many narrowband orthogonal carriers of 15 kHz each. Instead of a single fast transmission, a data stream is split into many slower data streams that are transmitted simultaneously. As a consequence, the attainable data rate compared to UMTS is similar in the same bandwidth but the multipath effect is greatly reduced because of the longer symbol duration [11].

If less than 5 MHz bandwidth is available, LTE can easily adapt and the number of narrowband carriers is simply reduced. Several bandwidths have been specified for LTE: from 1.25 MHz up to 20 MHz. The channel bandwidth used in practice depends on the amount of spectrum available to the network operator. In a 20 MHz band, data rates beyond 100 Mbit/s can be achieved under optimal signal conditions [11].

### 2.1.2 LTE Downlink Characteristics

This section lists a few of the LTE characteristics that are useful for the reader to understand the remaining of this work, or necessary to configure the VSA software settings:

- The standard LTE symbol duration is 66.7 µs. The corresponding orthogonal subcarriers spacing is 15 kHz.
- LTE supports several channel bandwidths. Furthermore, both frequency division duplexing (FDD) or time division duplexing (TDD) are supported. The experiments conducted in this research activity dealt exclusively with 10 MHz channels in FDD mode. However, the conclusions drawn in Chapter 5 are believed to be applicable to the other variants.
- Modulation types supported in the downlink are: binary phase shift keying (BPSK), QPSK, 16 QAM and 64 QAM. Zadoff-Chu sequences are also used for the primary synchronization signal (PSS) [12].
- The resource element is the smallest unit in the physical layer and occupies one 15 kHz subcarrier for one symbol duration. The smallest unit in resource allocation is however the resource block (RB), which occupies 12 adjacent subcarriers (180 kHz) of bandwidth during 7 symbols, or one slot [11].
- The frame structure is the same for the uplink and downlink transmission in LTE. However, the frame structure varies between FDD and TDD. The FDD frame structure is 10 ms-long and contains 20 slots of 7 symbols. An LTE subframe comprises two contiguous slots; there are therefore 10 subframes in an FDD frame, each 1 ms-long [11].
- LTE supports only packet-switched communication carried by shared channels. There are therefore no dedicated channels. LTE is the first cellular standard to rely exclusively on an packet-switched Internet protocol (IP)-based core network for both voice and data, with the exception of short messaging service (SMS), which is transported over signalling messages.

### 2.1.3 LTE Downlink Channels

LTE channels are defined logically. Logical channels do not occupy a specific subcarrier frequency. Instead, certain key channels are periodically transmitted in predefined RBs. The others are defined in a channel map that is transmitted and announces where specific logical channels are located in upcoming frames [13]. It is

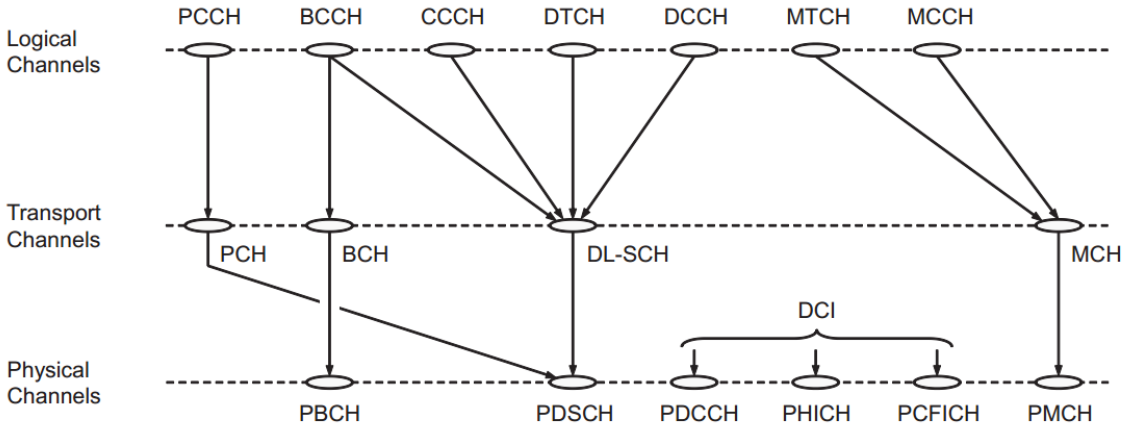



Figure 2.1: LTE downlink channel map, from [14].

important to note that logical channels are mapped to transport channels, which are in turn mapped to physical channels, as shown in Figure 2.1.

### Downlink Physical Channels and Signals

This section summarizes important downlink channels and reference signals. The UE needs to synchronize to the downlink signal before attempting to transmit and request to join the network [13].

**PSS and SSS:** The primary synchronization signal (PSS) and secondary synchronization signal (SSS) are two types of synchronization signals that are designed to be detected by all types of UEs. They are transmitted twice per 10 ms radio frame. They occupy the central 62 subcarriers of the channel, which ensures cell search is standardized regardless of the channel bandwidth. The PSS and SSS help the UE derive the physical cell identity (PCI), a variant of which was used as class label in the SVM classifier.

**PBCH:** The physical broadcast channel (PBCH) carries the broadcast channel (BCH) transport channel, which contains cell-specific content and is used for all types of UEs. It is transmitted in the centre of the channel and occupies 6 RBs (72 subcarriers). It is transmitted using QPSK.

**PDSCH:** The physical downlink shared channel (PDSCH) carries the traffic data and is shared in time between multiple users. QPSK, 16 QAM, and 64 QAM modulations are supported on the PDSCH.

**PDCCH:** The physical downlink control channel (PDCCH) carries the channel allocation and control information. It is transmitted using QPSK.

### 2.1.4 Orthogonal Frequency Division Multiplexing (OFDM)

Orthogonal frequency division multiplexing is the multiplexing scheme chosen for the LTE downlink [13]. The general idea of the OFDM transmission technique is to split the total available bandwidth $B$ into many narrowband sub-channels at equidistant frequencies. The sub-channel spectra overlap each other but the subcarrier signals are still orthogonal. The single high-rate data stream is subdivided into many low-rate data streams for the sub-channels. Each sub-channel is modulated individually using a conventional modulation format such as QAM and is transmitted simultaneously in a superimposed and parallel form [10, 13].

OFDM has the ability to perform well through a low quality channel, is immune to frequency-selective fading and provides resistance to inter-symbol interference in a multipath environment by reducing the symbol rate transmitted on each subcarrier [13].

Since the system bandwidth $B$ is subdivided into $N$ narrowband sub-channels, the OFDM symbol duration $T_S$ is $N$ times larger than in the case of an alternative single carrier transmission system covering the same bandwidth $B$. Typically, for a given system bandwidth, the number of subcarriers is chosen in such a way that the symbol duration $T_S$ is sufficiently large compared to the maximum multi-path delay $\tau_{\max}$ of the radio channel [10].

Figure 2.2 shows the power spectral density of a single OFDM subcarrier, whereas Figure 2.3 shows the power spectral density of multiple OFDM subcarriers of constant amplitude. LTE supports several constant amplitude modulation schemes that would result in a power spectral density display as shown in Figure 2.3 but also supports variable amplitude modulation schemes such as 16 and 64 QAM [13].

Today, OFDM is widely used in applications ranging from digital television and audio broadcasting to wireless networking such as IEEE 802.11 and wired broadband Internet access [13, 15]. Although OFDM was considered as a candidate for GSM in the 1980s, and seriously considered again as a candidate for the UMTS standard, other multiplexing schemes were favoured due to the high cost of computing power. However, with today's availability of small, low-cost, low-power chipsets, OFDM has become the technology of choice for the next generation of cellular wireless networks.

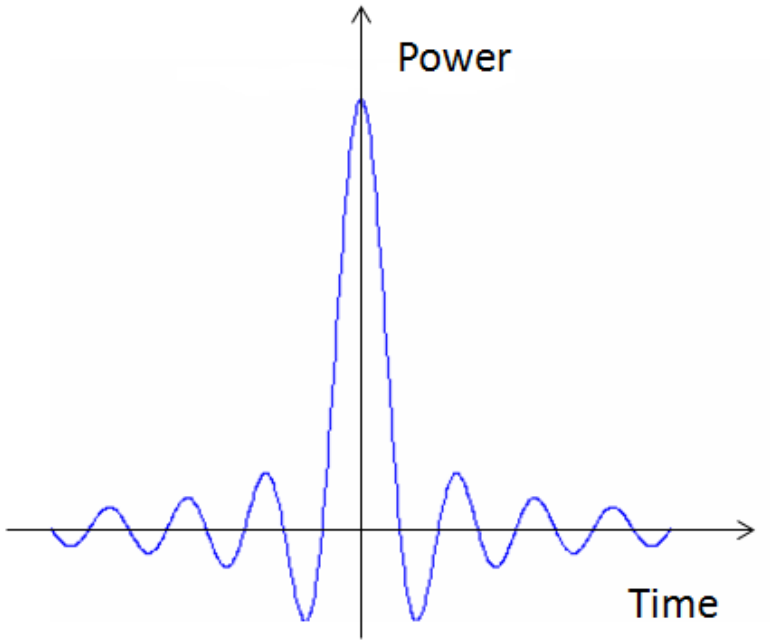



Figure 2.2: Spectrum of a single modulated OFDM subcarrier (truncated), from [13].

The first cellular system to adopt OFDM was IEEE 802.16e (mobile WiMAX). It was followed soon after by IEEE 802.20, the basis for Third Generation Partnership Project 2 (3GPP2)'s Ultra-Mobile Broadband (UMB), now abandoned, and most recently by the Third Generation Partnership Project (3GPP) for LTE [13].

Contrary to CDMA schemes widely used in 3G cellular systems, OFDM is able to perform frequency selective scheduling using real-time feedback of channel conditions. It is also completely free of multipath distortions up to the cyclic prefix, which is possible because of the long period of each OFDM symbol. These attributes of OFDM, in addition to the simpler equalization for large bandwidths and the better suitability to multiple inputs multiple outputs (MIMO) operations were key considerations for both working groups (3GPP and 3GPP2), as well as the IEEE 820.16 working group to use subcarrier-based transmission and multiplexing schemes such as OFDM and single-carrier frequency division multiplexing (SC-FDM) [13].

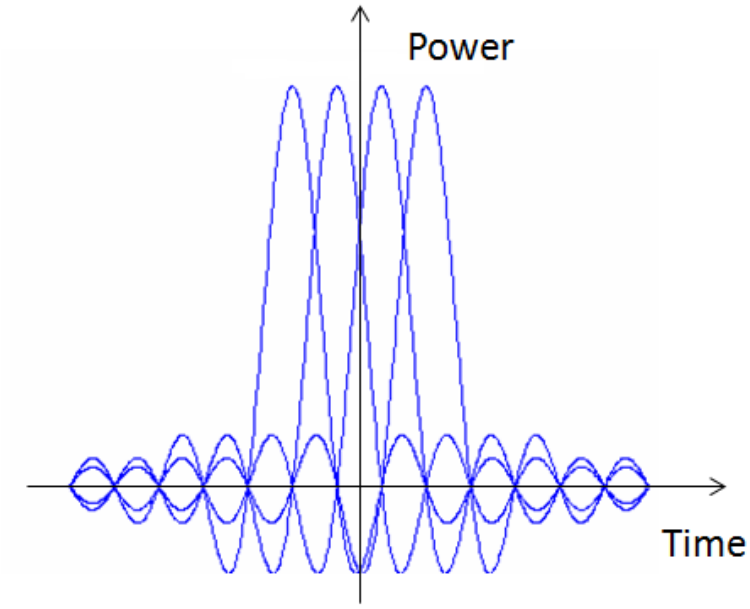



Figure 2.3: Spectrum of multiple OFDM subcarriers of constant amplitude, from [13].

### 2.1.5 Orthogonal Frequency Division Multiple Access

LTE uses a variant of OFDM in the downlink called orthogonal frequency division multiple access (OFDMA) [13]. In OFDM, a single user is allocated a number of subcarrier channels for the duration of the transmission, up to the entire set of subcarrier frequencies. In OFDMA however, users are allocated to a set of subcarrier channels for a single symbol duration. Other users are subsequently allocated the same subcarrier channels during the next symbol in time. From the perspective of a single user, his subcarrier channel allocation appears to be frequency hopping. The difference between OFDM and OFDMA is depicted in Figure 2.4.

## 2.2 Emitter Discrimination

The ability to uniquely identify communication and radar emitters is a concern for many defence, law enforcement and security applications. In communication systems, this requirement stems from the fact that wireless communication systems allow users access from untethered location, making authentication more difficult. Most communication systems rely on cryptographic security as well as a unique device identifier which is submitted to gain service access. However, the device identifier is often

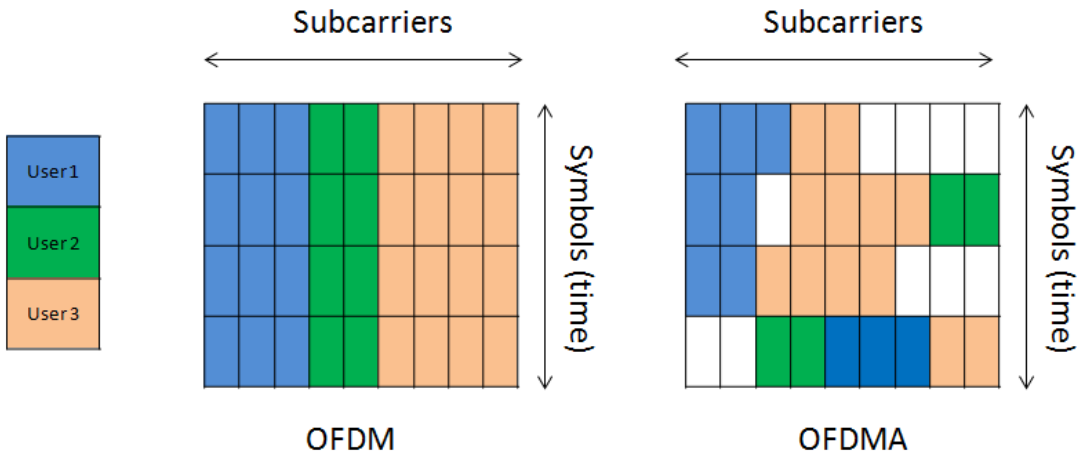



Figure 2.4: OFDM and OFDMA subcarrier allocation, from [13].

unreliable as they can be altered or copied [1].

In defence applications, the ability to uniquely identify a communication or radar emitter enables precise geolocation amongst many similar emitters sharing the same frequency channel, as well as tracking such emitter when it is in motion [16, 17]. Similarly, user identification based on imperfections in the transmitter RF chain can be used to recognize and identify criminals masking their digital identity by changing their device identifier [1]. Criminal activities who may gain from masking their wireless identities include sexual exploitation of children, production and illegal dissemination of copyright-protected media, intellectual property theft, identity theft, financial fraud and espionage [18].

Two main approaches to conduct radiometric identification emerged in the open literature. The first one consists in exploiting channel information to distinguish user location and detect when the same device identifier appears to be transmitting from different locations to detect forging. In a rich multipath environment, because of rapid path decorrelation, users can almost be uniquely characterized by their channel conditions [19]. Li et al. argue in [20] that in this situation, the response of the medium along any transmit-receive path is frequency-selective (or in the time domain, dispersive) in a way that is location-specific. Another channel-based fingerprinting technique observes received signal strength (RSS) values associated with packets measured at one or more receiver antennas. The RSS values are correlated with transmission power, the separation between the transmitter and the receiver,

and the complexity of the radio environment in which communication takes place [21]. However, these techniques are typically effective only in static settings, as it is well-known that RSS values can oscillate even in non-adversarial settings with legitimate users who are mobile. In such scenarios, identification of emitters using RSS and other physical layer parameter-based solutions relying on channel information result in a large number of false positives [21]. For these reasons, emitter discrimination based on channel information will not be discussed further.

The second approach focuses on hardware imperfections present in each transmitter rather than characteristics of the radio channel. These imperfections can be studied in the waveform domain or in the modulation domain [2]. In the waveform domain, most research focuses on the transient portion of a signal, whether it constitutes a symbol in a communication system or a pulse in a radar system. A transient is a brief radio emission produced while the power of the output of an RF transmitter goes from zero to the level required for the application to be effective, be it communications, or radar detection [1]. In transient-based communication systems, efforts are made towards characterizing the transient waveform at the beginning of each frame. In radar systems, the physical characterization of the unintended modulation of each pulse is sufficient to distinguish between radar emitters of the same class [22]. However, transient-based identification is more challenging for communication emitters: the low transmit power and short duration of the transient is difficult to detect, and describing the resulting waveform in a succinct way is challenging [1]. Waveform-domain steady-state signal characterization was also conducted by Kennedy et al. in [23], as well as Zamora et al. [24].

On the other hand, studies of hardware imperfections in the modulation domain generally consists of cataloguing selected features of a communication transmitter. These features can be extracted using commercial VSA software. This technique represents signals at the most basic level in terms of in-phase and quadrature (I&Q) samples, whose interpretation depends on the underlying modulation scheme. Signals in the modulation domain are more structured and better behaved, but require knowledge of the modulation scheme being used [2].

## 2.3 Hardware Imperfections

Due to imperfect manufacturing processes of RF components, minute differences are introduced by emitters during modulation, even for emitters of the same make and model. It is therefore possible to distinguish emitters by characterizing aspects of the transmitted signal. Since most of this work attributes the ability to conduct radiometric identification and radio-fingerprinting on hardware imperfections in the transmitters' RF chain, a discussion regarding the source of these imperfections is warranted.

Each component of the transmitter chain demonstrates imperfections caused by non-idealities of production processes. Metal-oxyde semiconductors, from which the components' circuits are made of, exhibit broad variations in major device parameters (e.g. channel length, channel doping concentration, oxide thickness) among production lots. These variations may occur for many reasons, such as minor changes in the humidity or temperature in the clean-room, or due to the position of the die relative to the centre of the wafer. Changes in device parameters influence transistors switching speed and thereby components' characteristics. Similarly, parameters of passive electronic devices follow random distributions caused by production inaccuracies, rather than taking a constant and uniform specified value. Despite technological advancements, constant market push for low-price, high-volume products results in variations among individual devices caused by the production imperfections. These variations, while being small enough to fulfil the requirements specified in the communication standards, are significant enough to allow for unique characterization of these devices via RF transceiverprints [1].

## 2.4 Radiometric Identification Process

Radiometric identification follows a similar process in the majority of open literature. A training period consists of capturing the transceiverprint of the emitter. Transceiverprint gathering requires signal acquisition with sufficient precision, and is followed by a feature extraction. It may be necessary to extract features from several frames or pulses during the training period in order to build a device model that will suitably represent typical emissions rather than a specific ones. Varying channel conditions could, for example, introduce some variance between each frame or pulse which may render re-identification difficult.

During the re-identification process, the signal acquisition and feature extraction is repeated for unknown devices, and the set of features is compared with known features using various decision algorithms, leading to a determination whether the device is recognized and provision of a confidence level [25]. Again during the re-identification, it may be necessary to extract features from several frames or pulses in order to enhance detection reliability [2].

### 2.4.1 Signal Acquisition

Signal acquisition challenges vary significantly in transient-based and modulation-based approaches. In both cases, it is necessary to capture the radio signals of wireless devices with sufficient precision. This is an important requirement given that devices' fingerprints at the physical layer are due to small impairments/variations in the devices' radio circuitry that could be easily lost if captured with inappropriate hardware [26]. Captures in the waveform domain generally require more sensitive equipment and are more subject to multipath and fading. Researchers operating in the modulation domain can often use the same receiver hardware that is originally engineered for the communication system, and demonstrate higher success rates [2].

### 2.4.2 Feature Extraction

The feature extraction process consists of extracting/selecting features from the radio signal that have sufficient discriminative capabilities to distinguish a given device and/or a class of devices [25].

Feature extraction is more complex in transient-based systems in which a waveform needs to be characterized. The technique employed by Shaw and Kinsner in [27] calculates the variance of the amplitude for each consecutive portion/window of the signal and compares each of these values, in sequence, to a predetermined threshold. The start of the transient is detected when the variance exceeds the threshold by a given margin. The end of the transient is determined in an experimental manner. The drawback of this approach is that the estimation of the threshold is difficult, given that the amplitude of the signal is susceptible to noise and interference. Another approach, which is also based on the variance of the amplitude, is the Bayesian step change detector (BSCD). The underlying technique, proposed by Ureten and Serinken in [28], transforms a change in the variance into a change in the mean value

that is subsequently used by the BSCD to detect the start of the transient. Unlike the previous approach, the detection of the transient is based exclusively on the characteristics of the amplitude data. Consequently, this approach can be used with various types of signals. However, the performance is less than optimal for signals that exhibit a gradual change in power at the start of the transient, such as IEEE 802.11 and Bluetooth. The same authors have recently proposed in [29] an enhanced detection method, referred to as the Bayesian ramp change detector to accommodate these signals [6].

In [4], Hall et al. chose to extract features from all three main components of the transient: amplitude, phase and frequency. The feature vectors include the standard deviation of normalized amplitude, the standard deviation of normalized phase, the standard deviation of normalized frequency, the variance of change in amplitude, the standard deviation of normalized in-phase data, the standard deviation of normalized quadrature data, the standard deviation of normalized amplitude (mean-centered), the power per section, the standard deviation of phase (normalized using mean) and the average change in discrete wavelet transforms coefficients using the Daubechies filter.

In modulation-based identification, features are usually extracted using test equipment. In [2], Brik et al. chose five features output by the VSA software which gave best re-identification results. Brik et al. extracted five features from each transmitted frame, namely frequency error, SYNC correlation, I&Q origin offset, error vector magnitude and error vector phase [25].

### 2.4.3 Decision-Making Algorithms

Several algorithms can be used to find the best match between databased features, gathered during the training period, and those of an unknown emitter gathered during the identification period. In [2], Brik et al. demonstrated the $k$-nearest neighbours (kNN) as well as the SVM algorithms, with increased accuracy using the SVM.

Liu et al. compare two new online clustering algorithms that are developed for radar emitter classification in [16]: one based on the minimum description length (MDL) criterion and the other on competitive learning. The model-based algorithm is shown to surpass the competitive learning algorithm in terms of classification accuracy, flexibility, and stability [16].

Probabilistic neural networks (PNN) have also been used by many research teams,

however the issue of scalability (memory requirement per profile) prohibits its use in real time systems [4]. Instead, Hall et al. favoured a statistical classifier which uses a set of features to represent a vector that is to be classified. The probability of a match is calculated using a modified Kalman filter from Bar-Shalom.

In [18], Dolatshahi et al. consider the performance of the generalized likelihood ratio test (GLRT) and a classical likelihood ratio test to match emitters based on power amplifiers characteristics, which features are extracted using Volterra series, with good results. GLRT outperformed the classical likelihood ratio test in most cases.

## 2.5 Transient and Steady-State Signals

Radiometric identification of emitters using waveform characterization, particularly during the transient period, has the longest history and is still the only method by which radars can be discriminated.

### 2.5.1 Radar Emitters

In radar applications, for example, it is not difficult to distinguish radar systems which transmit pulses of different radio frequencies or pulse repetition interval (PRI). However, in modern radar systems, more sophisticated signal waveforms are used and inter-pulse information may not be enough to separate those received pulses according to their originations. To classify radar emitters in such an environment, the detailed structure inside each pulse, called intra-pulse information, needs to be examined and characterized. This is because emitters exhibit their own electrical signal structure inside each transmitted pulse, due to both intentional and unintentional modulations. The SEI is a composite task that involves pulse measurements, features extraction, normalization, selection, classification (recognition) and verification [30].

### 2.5.2 Communication Emitters

Hall et al. published a landmark paper concerning the radiometric identification of IEEE 802.11b communication emitters in [4]. The feature extraction and decision algorithms employed by Hall et al. are presented in Section 2.4.2 and Section 2.4.3. The use of a Bayesian filter, to probabilistically estimate the state of a system from

noisy observations, mitigated the increased variability between signals from a given transceiver due to interference. The experiment consisted of collecting 100 signals from each of the fourteen 802.11b transceivers. Results achieved using transient-based radio frequency fingerprinting and a Bayesian filter neared 100%.

### 2.5.3 Volterra Series Coefficients

In [1], Polak et al. examine the non-linearities of two components of the transmitter RF chain, namely the digital-to-analog converter (DAC) and the power amplifier, with a view to unmask the identity of criminal users. The DAC's integral non-linearity specifies the actual output level for a given input word, from the ideal output level, and is caused by production inaccuracies that cause output levels of individual analogue sources from the DAC to vary around their nominal values. Secondly, power amplifiers are attractive for digital forensics purposes in that they are the last elements of the transmitter chain and are therefore the most difficult for a user to modify via software or even baseband control. In [1], the non-linear characteristics of power amplifiers were modelled using Volterra series representations.

In contrast with the work of Hall et al. in [4], and the work of Brik et al. in [2], the work of Polak et al. provides a complete statistics-based model which considers the effect of signal-to-noise ratio on the probability of successful emitter recognition.

### 2.5.4 Universal Mobile Telecommunications System

Kennedy and Kuzminskiy present in [31] a reliable way to uniquely distinguish between UMTS transmitters using steady-state characterization in the waveform domain. The proposed algorithm differs from previous work by performing joint channel estimation and classification on a steady state signal. The technique may be applied to any radio system with a repeated symbol sequence - such as a preamble. The laboratory demonstrator presented is capable of distinguishing between the preamble signal transmitted by UMTS UEs. Excellent results - in excess of 99% for 20 different UMTS models in an indoor wireless environment are reported. Kennedy and Kuzminskiy also comment on the work of Brik et al. in [2] as the frequency offset error between transmitter and receiver dominates the discriminatory performance of their solution. Whilst frequency offset applies to 802.11, it is not easily applicable to UMTS and other systems where the handset constantly disciplines its local timing

source to the base station broadcast channels. The much tighter front-end specifications for frequency offset and error vector magnitude of UMTS handsets also make their discrimination a more difficult problem than the 802.11 case. The raw measurement data was first collected using a Rohde and Schwarz FSQ26 signal analyser, then processed in Matlab. The processing stage involved locating and extracting the preamble bursts and filtering out any preambles of lower power that may have been captured, from multipath. The coherency algorithm takes the training file, compiled in an anechoic chamber, and the wireless test bursts to produce the confusion matrix.

## 2.6 Modulation Domain

Based on the higher identification accuracy of modulation-based transceiverprints, greater attention was devoted in reviewing publications that selected this approach. The first two papers reviewed below both performed transceiver fingerprinting of IEEE 802.11b transmitters, a wireless data standard that uses OFDM at the physical layer. Research in modulation-based radiometric identification of 4G cellular technologies, such as LTE, were not found. The most relevant OFDM work is thus reviewed next.

### 2.6.1 PARADIS

In [2], Brik et al. proposed the Passive RAdiometric Device Identification System (PARADIS). The team conducted a modulation-based radiometric identification using five features provided by the VSA software, namely frequency error, SYNC correlation, in-phase and quadrature (IQ) origin offset, symbol magnitude error and symbol phase error. Brik et al. achieved remarkably low radiometric identification error rates: using 138 unique network cards, from the same vendor and of the same model, they achieved an error rate of 0.34% using PARADIS with SVM classification. Invalid frames were discarded prior to the evaluation, that is frames with an invalid checksum or those that did not comply with the 802.11b standard tolerance for frequency and modulation error. Invalid frames accounted for 4% of the collected data.

### 2.6.2 Multiple Inputs Multiple Outputs Transmitters

Shi and Jensen present interesting results in conducting radiometric identification of MIMO transmitters in [32]. The feature selection aspect of this research is unique and discussed fully in Section 2.9. Shi and Jensen were able to obtain excellent results with both the bagging with the decision tree (DT) kernel and SVM (in excess of 99%). They also showed that MIMO classification provides increased robustness in face of high feature variance, due to the enriched feature set.

## 2.7 Reliability of Radiometric Identification

RF chain imperfections cannot be altered by the user without significant effort [1]. However, it may be possible to reproduce another user's identity by mimicking its RF imperfections using high-end signal generators. Danev et al. attempted, and mostly succeeded in defeating radiometric-based identification systems in [25]. In spite of the accepted belief that hardware imperfections are hard to reproduce, different impersonation attacks are proposed and demonstrated. It is shown that modulation-based identification can be impersonated with accuracy close to 100%. However, successful attacks require specialized hardware and often also require a knowledge of the features that are examined and the classification algorithm. Indeed, the success rate of the impersonation highly depends on the discriminant capabilities of the classifier used [25]. The SYNC correlation, in particular, could not be accurately reproduced with the software defined radio platform, and is also listed as one of the most reliable criteria by Birk et al. in [2], thus a classifier which would rely heavily on this feature would be harder to fool. In related work by Edman and Yener in [33], notable success (75%) was achieved in fooling an 802.11 radiometric identification system using the universal software radio peripheral (USRP), a commodity software-defined radio system. However, success rates only slightly above 50% were achieved when injecting simulated packets rather than replaying sampled recordings. These attacks were also performed with a-priori knowledge of the recognition features.

Transient-based features could also be reproduced with a high-end arbitrary waveform generator over a wire, but the attacks were largely unsuccessful over the wireless medium. This is due to channel multipath effects on the transient part of the signal. Feature extraction at the receiver will differ from those of the attacker trying to clone because the channel effects will vary between the pairs attacker-receiver and

transmitter-receiver [25].

The complexity involved in conducting a successful impersonation attack against a system which discriminates based on physical-layer behaviour is two-fold. 1) The attacker requires specialized hardware to capture the target device's signal, not unlike the one of the system under attack. 2) The attacker also requires the specialized hardware needed to reproduce features of this signal with sufficient fidelity to fool the security system. Danev et al. conclude in [25] that physical-layer identification is, alone, not a suitable technique to enforce access policy and that other authentication techniques should be used in parallel. This echoes the assessment of Brik et al. in [2].

Polak and Goeckel investigated in [19] the radiometric identification of users actively trying to falsify their RF signatures by injecting slight distortion to the data symbols. In simulations based on parameters of commercially-used power amplifiers, it is shown that the transmitter identification performance of modifying data symbols is similar to users who are not.

## 2.8 Support Vector Machines

The work of Brik et al. in [2] indicate that SVMs are well suited for radiometric identification problems, and that SVMs outperform other classifiers. Other classifiers, such as the PNN discussed by Hall et al. in [4], presented some scalability problems rendering it difficult to use in real-time systems. As a result, the decision was made to use the SVM classifier to validate the hypotheses. Comparing re-identification accuracy with other classifiers was not a research objective for this thesis, and as such is left to Section 5.3 (Future Work). An SVM is an abstract learning machine which will learn from a training dataset and attempt to generalize and make correct predictions on novel data. For the training data, we have a set of input vectors, denoted $\mathbf{x}_i$, with each input vector having a number $m$ of component features. These input vectors are paired with corresponding labels, which we denote $y_i$, and there are $n$ such pairs ($i = 1, \ldots, n$) [34]. The model built during training will be used to predict the correct label $y$ for a given test vector $\mathbf{x}$.

For two classes of well separated data, the learning task amounts to finding a directed hyperplane, that is, an oriented hyperplane such that data points on one side are labelled $y_i = +1$ and those on the other side as $y_i = -1$. The directed hyperplane found by a SVM is intuitive: it is that hyperplane which is maximally

distant from the two classes of labelled points located on each side. The closest such points on both sides have most influence on the position of this separating hyperplane and are therefore called support vectors [34].

The separating hyperplane is given as $\mathbf{w}\cdot\mathbf{x}+b=0$, in which $\cdot$ denotes the inner or scalar product, $b$ is the bias or offset of the hyperplane from the origin in input space, and $\mathbf{x}$ are points located within the hyperplane. The normal to the hyperplane, the weight vector $\mathbf{w}$, determine its orientation [34].

The SVM optimization expression is presented in Equation (2.1)

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to:} \quad & y_i\left(\mathbf{w}^T\phi\left(\mathbf{x}_i\right)+b\right) \geq 1-\xi_i, \\ & \xi_i \geq 0 \end{aligned} \tag{2.1}$$

given a training set of $n$ instance-label pairs $(\mathbf{x}_i, y_i)$, in which the feature vector $\mathbf{x}_i \in \mathbf{R}^m$, $y_i$ represents the class label, and where $\xi_i$ is a small number, and $^T$ is the transpose operator. $\Phi(\cdot)$ is a mapping function allowing data points to be mapped into a space with a different dimensionality, called *feature space*, with the replacement $\mathbf{x}_i\cdot\mathbf{x}_j \rightarrow \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$. $C>0$ is the penalty parameter of the error term, also termed regularization parameter [34–36].

As an example Figure 2.5 shows two labelled clusters which are readily separable by a linear hyperplane. In reality, the two clusters could be highly intermeshed with overlapping data points: the dataset is then not linearly separable, as shown in Figure 2.6. Using the mapping function $\Phi(\cdot)$, which maps the *input space* into a *feature space* of higher dimensionality, it becomes possible to find separating hyperplanes. In Figure 2.6, which does not appear to have a linear hyperplane, if one pushes the points associated with one class down into a third dimension, a hyperplane could be readily constructed parallel to the page in order to separate the two classes [34]. The hyperplanes generated by a given kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ result in non-linear hyperplanes of various shapes once represented in the *input space*. Many kernel functions are possible, each offering different tuning parameters: polynomial, Gaussian, sigmoid, feedforward neural network or radial basis function (RBF). In fact, kernel substitution can be applied to a wide range of data analysis methods so that SVMs should really be viewed as a sub-instance of a much broader class of *kernel-based*

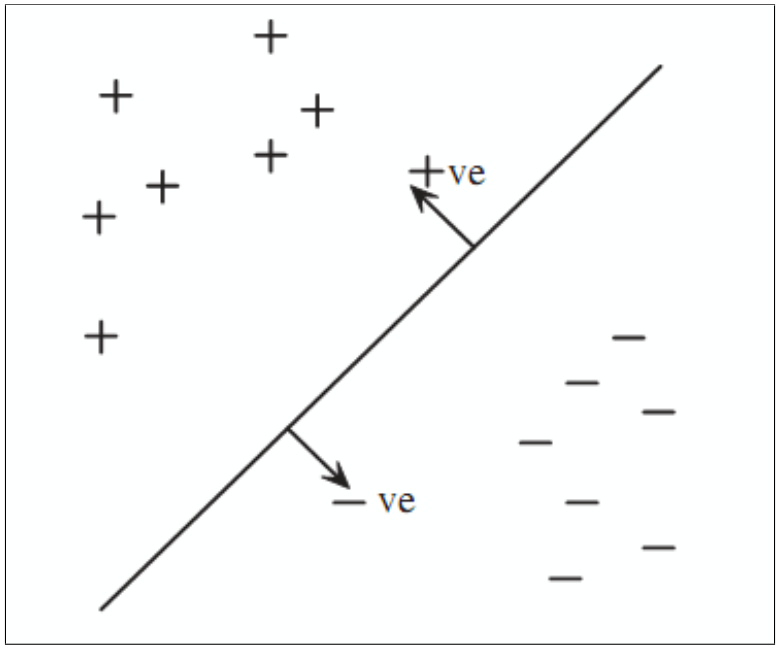



Figure 2.5: Linear hyperplane in SVM, from [34].

*methods* [34].

In general, the RBF kernel is a reasonable first choice. This kernel nonlinearly maps data points into a higher dimensional space so that cases when the relation between class labels and attributes is nonlinear can be handled. Furthermore, the linear kernel is a special case of RBF since the linear kernel with a penalty parameter $C$ has the same performance as the RBF kernel with some parameters ($C$,$\gamma$). In addition, the sigmoid kernel behaves like RBF for certain parameters. The second reason why RBF kernel is a reasonable first choice revolves around the number of hyperparameters, which in turn influences the complexity of the model selection. The polynomial kernel has more hyperparameters than the RBF kernel. Lastly, the RBF kernel has fewer numerical difficulties than kernels such as sigmoid or polynomial [35], simplifying its use. Equation (2.2) shows the definition of the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ as well as the mathematical expression for the RBF kernel. $\gamma$ is the only kernel parameter for the RBF kernel function.

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &\equiv \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \\ K(\mathbf{x}_i, \mathbf{x}_j) &= \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \gamma > 0 \end{aligned} \tag{2.2}$$

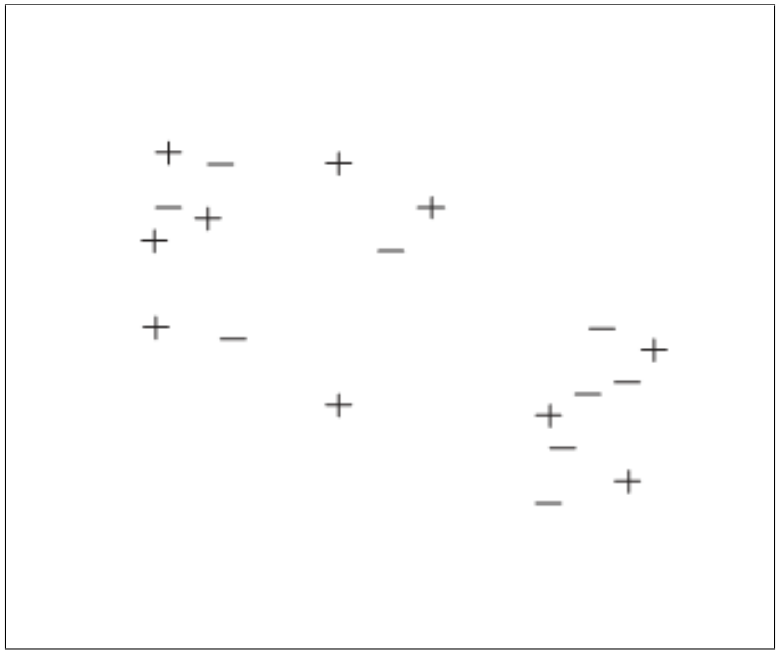


Figure 2.6: SVM hyperplane in a higher order dimension, from [34].

Parameters $C$ and $\gamma$ are further discussed in Chapter 3.

### Multi-class SVM

The problem in this thesis involves multi-class classification: each LTE transmitter's identity is used as a class label, and the SVM attempts to correctly assign unknown feature vectors to the best matching class. A number of schemes have been devised to extend binary SVM into multi-class classifiers:

- directed acyclic graph (DAG). If the number of classes is small then we can use a DAG with the learning task reduced to binary classification at each node. Suppose we consider a 3-class classification problem. The first node is a classifier making the binary decision, class 1 versus class 3. Depending on the outcome of this decision, the next steps are the decisions class 1 versus class 2 or class 2 versus class 3 [34].

- A series of *one-against-all* classifiers. We construct $N$ separate SVMs with the $n^{\text{th}}$ SVM trained using data from class $n$ as the positively labelled samples and the remaining classes as the negatively labelled samples. Associated with the $n^{\text{th}}$ SVM we have $f_n(\mathbf{z}) = \sum_i y_i^n \alpha_i^n K(\mathbf{z}, \mathbf{x}_i) + b^n$ and the novel input $\mathbf{z}$ is assigned

to the class $n$ such that $f_n(\mathbf{z})$ is largest. Though a popular approach to multi-class SVM training, this method has some drawbacks. For example, suppose there are 100 classes with the same number of samples within each class. The $N$ separate classifiers would each be trained with 99% of the examples in the negatively labelled class and 1% in the positively labelled class: these are very imbalanced datasets, and the multi-class classifier would not work well unless this imbalance is addressed [34].

- *One-class classification.* The idea is to construct a boundary around the normal data such that a novel point falls outside the boundary and is thus classified as abnormal. The normal data is used to derive an expression $\phi$ which is positive inside the boundary and negative outside. *One-class classification* is frequently used for novelty detection. *One-class classifiers* can be readily adapted to multi-class classification. Thus we can train *one-class classifiers* for each class $n$, and the relative ratio of $\phi_n$ gives the relative confidence that a novel input belongs to a particular class [34].

- *One-against-one.* If $N$ is the number of classes, then $N(N-1)/2$ classifiers are constructed and each one trains data from two classes. The classification uses a voting strategy: each binary classification is considered to be a voting, where votes can be cast for all data points $\mathbf{x}$ - in the end a point is designated to be in a class with the maximum number of votes. Hsu and Lin give in [37] a detailed comparison and conclude that *one-against-one* is a competitive approach. *libsvm* implements the *one-against-one* approach for multi-class classification [38].

## 2.9 Formal Feature Selection

Whereas Birk et al. empirically determined which of the five features leads to improved detection accuracy in [2], Shi and Jensen provide a formal method to choosing the VSA features based on maximum relevance and minimum redundancy. The latter team determine the best classifying features for a variable number of transmitters using the minimum-redundancy-maximum-relevance (mRMR) algorithm on the training dataset, where relevance is measured by the mutual information of that feature and the device identifier $h$:

$$I\left(g_i, h\right) = \int \sum_{h \in \mathbf{h}} p\left(g_i, h\right) \log \frac{p\left(g_i, h\right)}{p\left(g_i\right) p\left(h\right)} dg_i \tag{2.3}$$

where $\mathbf{h}$ represents the set of $N_d$ possible values of $h$, and $p\left(\cdot\right)$ denotes the probability density function. The relevance of a set of selected features $\mathcal{S} = \left\{g_1, g_2, \cdots, g_{N_f}\right\}$ is defined as the average of the relevance of features in the set, or

$$V_{\mathcal{S}} = \frac{1}{N_f} \sum_{g_i \in \mathcal{S}} I\left(g_i, h\right) \tag{2.4}$$

Similarly, redundancy is measured by the mutual information of two different features. The redundancy of the feature set is defined as the average of the mutual information of each pair of features in the set, or

$$W_{\mathcal{S}} = \frac{1}{N_f^2} \sum_{g_i, g_j \in \mathcal{S}} I\left(g_i, g_j\right) \tag{2.5}$$

Prior work on this topic has shown that finding the feature set that maximizes the difference $V_{\mathcal{S}} - W_{\mathcal{S}}$ is one simple approach that achieves the objective [32].

The result of their analysis is shown at Table 2.1. Interestingly, frequency offset is selected at most once even in cases with multiple transmitters. This is explained by Shi and Jensen because of the likelihood that all transmitter chains use the same local oscillator and as such the frequency offset for other transmitters is highly redundant [32].

Chen and Lin examine in [39] several feature selection algorithms combined with SVMs, namely F-score, random forests and radius margin (RM)-bound SVM, and compared against results obtained by SVM without feature selection. In the applications studied, the datasets contain a very larger number of features (500 to 100 000) and feature selection reduced the number of features entering the SVM model. They conclude that for some problems, such as optical character recognition, the use of the SVM classifier without feature selection performs well. In other problems, greater prediction accuracy is achieved with feature selection.

In extracting features of radar pulses, Kawalek and Owczarek discuss in [40] the set of intrapulse parameters: the rise, slope and fall time, rise angle, fall angle, angle of pulse, pulse point, frequency waveform, angle of frequency modulation, and the regression's line. Out of these parameters, a subset is selected to reduce computational

Table 2.1: Ordered significance of the features as a function of the number of transmitters $N_{TX}$ (subscripts indicate transmitter number), from [32].

| Rank | Number of Transmitters ($N_{TX}$) | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | Freq Offset | Freq Offset$_1$ | Freq Offset$_1$ |
| 2 | SYNC Corr | SYNC Corr$_1$ | SYNC Corr$_1$ |
| 3 | I/Q Offset | I/Q Offset$_1$ | I/Q Offset$_1$ |
| 4 | EVM | I/Q Offset$_2$ | I/Q Offset$_2$ |
| 5 | Symbol Clock Error | EVM$_1$ | I/Q Offset$_3$ |
| 6 | I/Q Gain Imbalance | SYNC Corr$_2$ | EVM$_1$ |
| 7 | I/Q Rotation | EVM$_2$ | EVM$_3$ |

complexity as well as minimize redundancy. Two techniques: the linear discriminant analysis (LDA) and the Karhunen-Loève transformation (KLT) were compared in order to determine this minimal set of parameters and provided near identical results.

In this work, the F-score and mRMR scoring algorithms are selected. F-score is simple and effective, and an implementation provided by Chen [41] is directly compatible with *libsvm* libraries and data formats. mRMR was also selected due its properties of reducing redundancy between features, and in order to compare its ranking with F-score.

## 2.10 Concluding Remarks

We presented an overview of several publications related to the radiometric identification of emitters, using a variety of feature extraction approaches, feature selection techniques and a multitude of classification algorithms. The work by Kennedy and Kuzminskiy, presented in [31], is closely related to this research effort and highlights some of the problems that can be forecast with 4G cellular technologies. The formal feature selection methodology of Shi and Jensen in [32] will be considered for the selection of LTE features. Note that at the time of writing, there are no academic publications discussing LTE transmitter identification using radiometric properties.

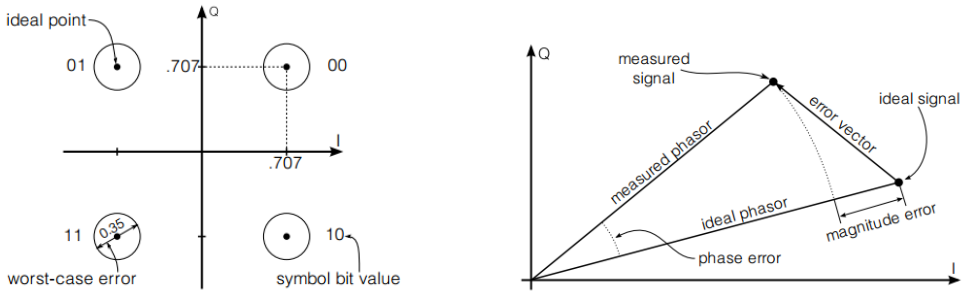
# Chapter 3

# Models and Theory

This chapter first outlines the LTE radiometric properties considered for this study. Then, details concerning the feature extraction and analysis process are presented. Mathematical models for the two feature ranking and five weighting algorithms are presented next. A discussion concerning signal quality filtering, aiming to improve re-identification accuracy, follows. Finally, an overview of the selection of a performance criteria concludes the chapter.

## 3.1 Radiometric Identification Features

The VSA software provides several signal traces for study; however, the experiments conducted in this thesis are centered on the *Error Summary* table. This decision was based on earlier published research by Brik et al. in [2] and the work of Shi and Jensen in [32], who also used a subset of the features available on the trace of the *Error Summary* table in their respective publications. It is also understood that traces which quantify RF and modulation errors are more likely to provide recognizable features originating from hardware imperfections in the RF circuitry of the transmitter. Chapter 5 presents other traces which could be studied in future research, which may also lead to transmitter-specific information useful for re-identification. Figure 3.1 provides a graphical representation of modulation errors and the calculation of the EVM, which are important notions for the remainder of this work. The *Error Summary* table provides the following values computed by the VSA software during the analysis and demodulation of LTE signals.




(a) Modulation errors (b) error vector magnitude

Figure 3.1: Graphical representation of modulation errors and error vector magnitude, from [2].

### 3.1.1 Error Vector Magnitude (EVM)

The EVM feature represents the root mean square (RMS) average of the error vector magnitudes for all channels and signals selected in the *composite include list* of channels, except non-allocated, over all symbol-times in the measurement interval. EVM can also be shown in units of decibel (dB) [42].

### 3.1.2 EVM Peak

EVM peak represents the signal's peak EVM value and location (symbol and subcarrier) within the frame. Only channels and signals that are selected for analysis in the *composite include list* of channels, except non-allocated, are included in the EVM. EVM peak value can also be shown in units of dB [42]. Although EVM peak value was considered for re-identification, the symbol and subcarrier indices were not.

### 3.1.3 Reference Signal EVM (RSEVM)

The RSEVM represents the RMS average of the reference signal (RS) error vector magnitudes for all symbol-times in the measurement interval. The value is displayed in units of % rms. RSEVM is calculated from the reference input channel for downlink signals [42].

### 3.1.4 Reference Signal Transmit Power (RSTP)

The RS transmit power (Avg) is the average reference signal power on the reference input channel for the data in the measurement interval. This data result can be used

to calculate RSTP as defined in 3GPP TS 36.141 Section F.3.3 [43]. This data result is only applicable to downlink signals [42].

### 3.1.5 OFDM Symbol Transmit Power (OSTP)

The OFDM symbol transmit power is calculated by averaging the power of all resource elements in the fourth symbol in every subframe. This is done for all subframes in the measurement interval and over all antenna ports, as defined in 3GPP Technical Specification 36.141, Section F.3.3 [43]. This data result is only applicable to downlink signals. The fourth symbol in every subframe contains only PDSCH, so this data result can be interpreted as the average power of the data subcarriers. OSTP is calculated and expressed as an absolute power in units of dB [42].

### 3.1.6 Reference Signal Received Power (RSRP)

The RS receive power (Avg) is a linear average of the power of all cell-specific RS resource elements from Tx antenna port 0 (and Tx port 1, when present) over the data in the measurement interval. RSRP (Avg) can be used to calculate RSRP as defined in Section 5.1.1 of 3GPP TS 36.214 [42, 43].

When there are multiple Rx input channels present that contain transmissions from Tx antenna port 0 or 1, the average RS power is calculated for each Rx input channel and added together to compute RS Rx Power (Avg) [42].

### 3.1.7 Reference Signal Receive Quality (RSRQ)

The RSRQ is a measure of the quality of the received signal as defined in Section 5.1.3 of 3GPP TS 36.214 and is given by the following equation:

RSRQ = $N$ * RSRP / (E-UTRA RSSI)

where:

$N$ is the number of resource blocks over which RSRP and evolved UMTS terrestrial radio access (E-UTRA) received signal strength indication (RSSI) are calculated (the LTE demodulator includes all resource blocks in the frame bandwidth). RSRQ is calculated over all data in the measurement interval [42, 43].

### 3.1.8 Frequency Error

The frequency error value represents the carrier frequency error relative to the VSA's centre frequency. This parameter is displayed in Hz and is the amount of frequency shift, from the VSA's centre frequency, that the VSA must perform to achieve carrier lock. In the downlink, for the fast Fourier transform (FFT) demodulation, the carrier frequency is initially determined from the preamble. The frequency error reported in the summary table is the value from the final frequency estimation based on the channel estimation sequence, with a fine adjustment made for any frequency error detected in the pilot tracking. The pilot tracking does not correct the signal for frequency errors, it only corrects the post-FFT symbols for phase (and amplitude, if selected) [42].

### 3.1.9 SYNC Correlation

SYNC correlation is the correlation coefficient between the measured preamble and an ideal preamble. This can be used as an indication of the quality of the preamble. A value of 1 indicates perfect correlation and a value of 0 indicates no correlation.

Large frequency errors may cause the VSA software to show incorrect low SYNC correlation values. These values are a result of frequency error and do not necessarily indicate poor signal quality. Therefore, for low SYNC correlation values you always need to validate the cause of the low SYNC correlation data result [42].

### 3.1.10 Common Tracking Error (CTE)

The CTE is the RMS average of the symbol-by-symbol deviation from the channel-compensated (equalized) signal expressed in % rms. An ideal signal would not have to be tracked (no magnitude scaling or phase adjustment) and the tracking adjustment value would be (magnitude 1, phase 0) for every symbol-time.

For each symbol-time in the CTE trace, the ideal signal's CTE magnitude is subtracted from the measured signal's CTE magnitude and the differences are averaged by RMS. Since the ideal signal's CTE magnitude is 1, the calculation can be simplified to the equation shown below.

$$\mathrm{CTE}_{\%\mathrm{rms}} = \sqrt{\sum_{t=0}^{N-1} \left(\left|\mathrm{CTE}_m\left(t\right)\right| - 1\right)^2} \times 100\% \tag{3.1}$$

where $\mathrm{CTE}_m\,(t)$ is the measured common tracking error at symbol-time $t$, and $N$ is the number of symbols in the measurement interval [42].

### 3.1.11 Symbol Clock Error

The symbol clock error shows the frequency error between the measured signal's symbol clock and the reference symbol clock in parts-per-million (ppm). The error is calculated by averaging the symbol timing corrections during the measurement interval. The symbol clock error is calculated from the reference input channel for downlink signals.

A value of 0.010 ppm indicates that the measured signal's symbol clock frequency is slower than the reference symbol clock frequency by $0.00000001 \times RefSymClk$ [42].

### 3.1.12 Time Offset

Shows the time offset from the beginning of the time capture to the beginning of the measurement interval.

For example, when time offset = 5 ms and the search time trace starts at −8.3 ms, the beginning of the measurement interval is located at −3.3 ms within the search time trace.

The time offset can be used to measure the accuracy of an external frame trigger signal when such signal is connected to the external trigger, the *analysis start boundary* is set to *frame*, and the *measurement offset* is set to 0. In this case, time offset will show the offset between the frame trigger and the actual start of the frame [42].

### 3.1.13 I&Q Offset

The I&Q offset indicates the magnitude of carrier feedthrough (power at 0 Hz). When there is no carrier feedthrough, I&Q offset is zero (-infinity dB). I&Q offset is calculated by computing the RMS average of the measured I&Q offset for each symbol in the measurement interval, and is expressed relative to the average signal power.

For a downlink signal, I&Q offset is calculated from the reference input channel. For an ideal downlink signal, any I&Q offset present would not affect the quality of the signal since the direct current (DC) subcarrier is orthogonal to the other subcarriers. However, when the signal is impaired so that the subcarriers are not orthogonal to

DC (when Doppler shifted, for instance), any I&Q offset on the DC subcarrier will affect the EVMs of other subcarriers [42].

## 3.2 Feature Extraction and Analysis Process

A lengthy feature extraction process was performed by the VSA software once the RF signals were recorded to disk during the field trial. The feature extraction relies on the VSA software's LTE demodulation analysis capability. The VSA software can display a number of analytical traces for either live signals or recorded ones. Using the VSA application programming interface (API), it is possible to automate much of this process. Specifically, with a program written in Visual Basic for Applications (VBA) (*VSA_Interface.xlsm*), it was possible to configure the settings of the VSA software to ensure repeatability, load signal recordings, and step through extracting trace values from loaded signal recordings to generate feature vectors. The feature extraction first populated a Microsoft Excel spreadsheet and was subsequently written to disk in the comma-separated values (CSV) format. The *Error Summary* table was selected for most of the experimentation and provides several measurements taken from the signal recording, such as EVM, SYNC correlation, symbol clock error, time offset, I&Q offset, I&Q gain imbalance, group and cell identification strings. From observations, it became also interesting to capture other VSA status indicators such as application synchronization status and raw synchronization status for every sample extracted. Another trace, the *DL Decode Info*, presented the decoded frame number, and this value was also used to enrich the feature extraction as one additional attribute of signal reception quality. Lastly, the trace extraction was enhanced with the sample time stamp. The file naming convention shown in Table 3.1 was used for all trace extracts.

A set of Linux shell scripts and simple programs were written to process the trace extracts and interface with external libraries for feature scoring and weighting, and in order to perform SVM analysis. The process flow is captured in Figure 3.2.

In all, near 10 000 SVM classification experiments were conducted. Cloud-based computing resources were used to conduct the analysis, using as many as three hosted Linux virtual private server (VPS) concurrently, in order to accelerate the process. The purpose and language of each software component is summarized in Table 3.2

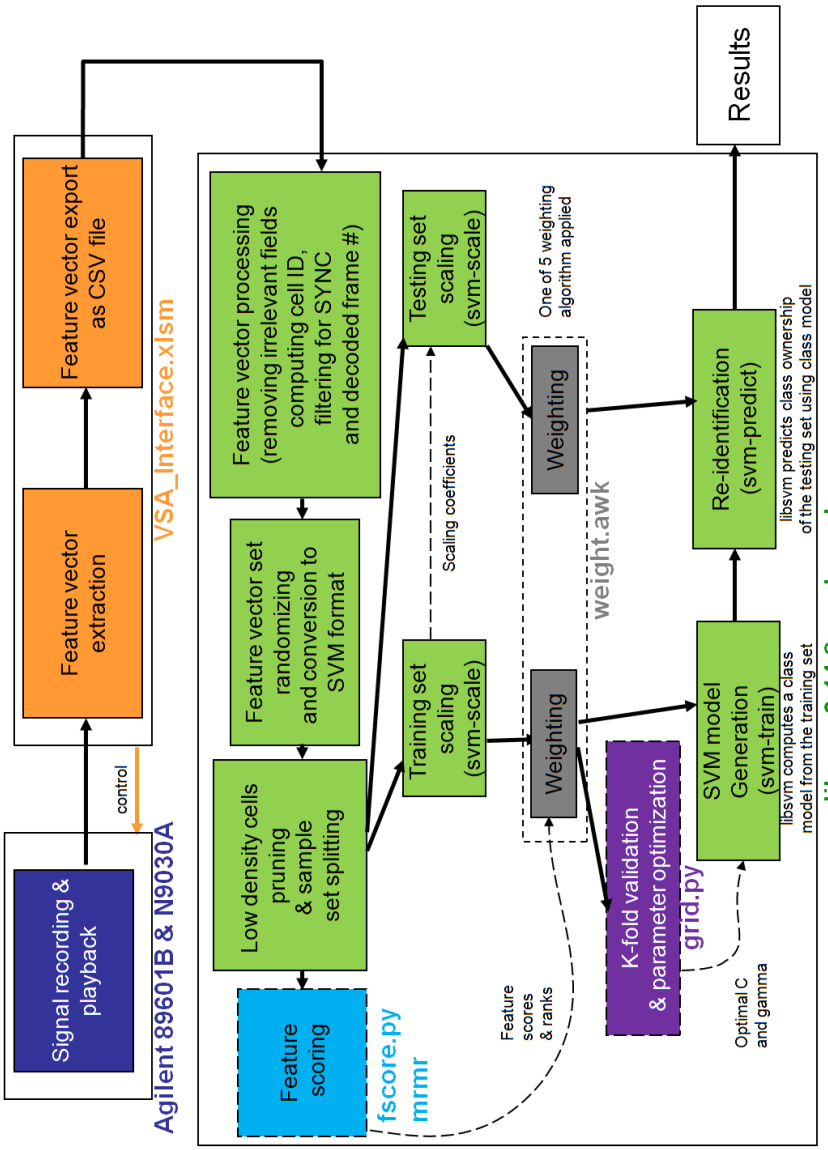



Figure 3.2: Analysis process.

Table 3.1: Trace extracts file naming convention.

| e.g. DL-10MHz-2150MHz-10s-13Feb12-c87-p001-n1000-o200-d0.5-t1.csv | | |
|---|---|---|
| **Example** | **Template** | **Description** |
| DL | UL/DL | uplink or downlink |
| 10MHz | xxMHz | signal bandwidth |
| 2150MHz | xxxxMHz | signal centre frequency |
| 10s | xxxs | recording duration |
| 13Feb12 | ddmmmyy | recording date |
| c87 | cxxx | forced cell ID (LTE Demod properties) (optional) |
| p001 | pxxx | recording position number (from map) (optional) |
| n1000 | nxxxxx | number of feature vectors extracted |
| o200 | oxxx | offset start of extraction (ms) |
| d0.5 | dxx.x | time span of each feature vector (ms) |
| t1 | tx | trace number (1:Error Summary Table) |

and detailed in the following subsections.

### 3.2.1 *VSA_Interface.xlsm*

The *VSA_Interface.xlsm* routines were used to automate the trace extraction from the Agilent VSA software. The tool written in VBA called upon the documented VSA API to first setup the VSA and the LTE demodulation settings from a pre-configured file, in addition to user-configurable parameters such as the trace of interest, the number of vectors to extract, the duration of each data point in ms (which needed to match the LTE demodulation parameters), as well as an offset from the beginning of the recording, also in ms. The user then selects the signal recording of choice and launches the trace extraction. The application provides a status indicator and steps through each feature extraction by moving the playback position and fetching the values of interest from the VSA trace. As an example, a 2000-point extraction, which was the usual case following the field trial in the City of Ottawa, lasted slightly over 20 minutes. Once the extraction finished, the feature vectors could be saved to file

Table 3.2: Analysis programs.

| **Name** | **Programming language** | **Description** |
|---|---|---|
| *VSA_Interface.xlsm* | VBA | Microsoft Excel spreadsheet responsible for the interaction with the Agilent VSA software. The spreadsheet VBA code automatically loads signal recordings and configuration settings in the VSA software, and steps through the RF recording, saving feature vectors for every frame. |
| *RecordingsIndex.accdb* | Microsoft Access SQL and VBA | Database containing all feature vectors and empirical observations taken during the field trial. The database VBA code was also used to compute basic statistics and generate a dynamic map of recording positions, as shown in Figure 3.6. |
| *analyze.sh* | Linux Bash | Core analysis program, used to process feature vector sets, and invoke external libraries. |
| *looper.sh* | Linux Bash | Simple script used to perform multiple experiments in an iterative fashion, without user intervention. |
| *fscore.py* | Python | Feature scoring using the F-score algorithm, from Chen [41]. |
| *mrmr* | C/C++ | Feature scoring using the mRMR algorithm, from Peng et al. [44]. |
| grid.py | Python | Utility provided with *libsvm* which performs a grid search to optimize SVM parameters $C$ and $\gamma$. |
| weight.awk | awk | Script used to apply feature weights using one of five weighting algorithms. |

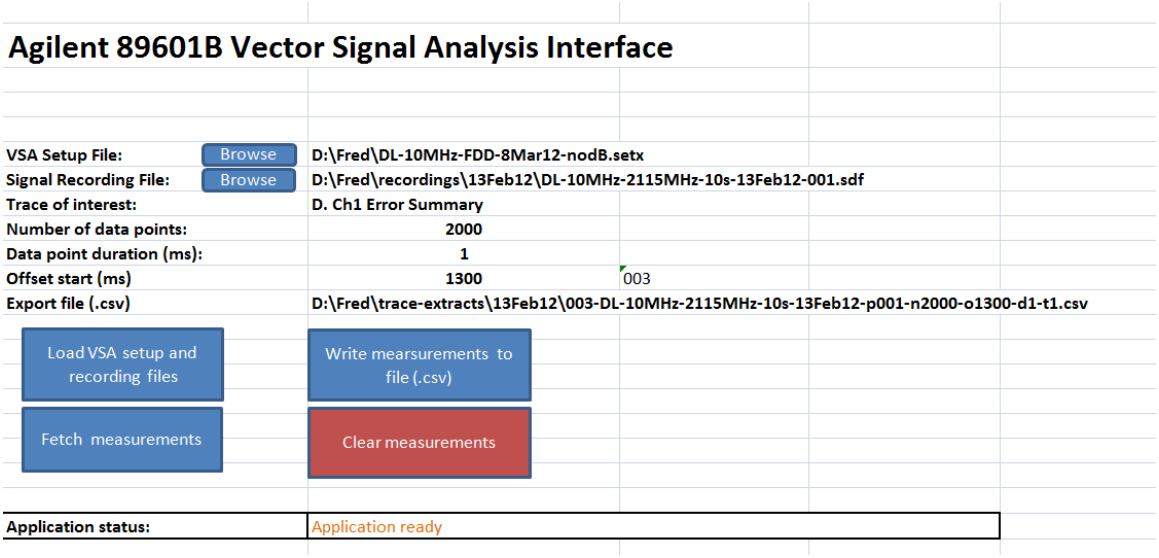



Figure 3.3: Screen capture of *VSA_Interface.xlsm*.

containing CSV. A screen capture of the *VSA_Interface.xlsm* is shown in Figure 3.3. A sample dataset output from *VSA_Interface.xlsm* is shown in Figure 3.4.

### 3.2.2 *RecordingsIndex.accdb*

All trace extracts were imported into a Microsoft Access database from which a number of utilities were created. A Microsoft Access form is used to show and retrieve the catalogue of recordings for any of the 36 recording positions, and provides useful information concerning the quality of the signal observed, the number of feature vectors, the number of feature vectors with synchronization (as reported by the VSA API) as well as the number of feature vectors with both synchronization and a decoded LTE frame number. Empirical observations taken during the recordings (such as signal strength) were also entered in the database. Queries built with the structured query language (SQL) into the Access form provide a list of most frequent cell IDs observed within each RF recording. A separate query form was added which facilitates the search for a specific cell ID, returning a list of recordings and positions where the cell ID was observed and recorded. A screen capture of the Microsoft Access database is shown in Figure 3.5.

In addition, a VBA module was created within the database in order to generate

Figure 3.4: VSA trace extraction showing feature vectors.

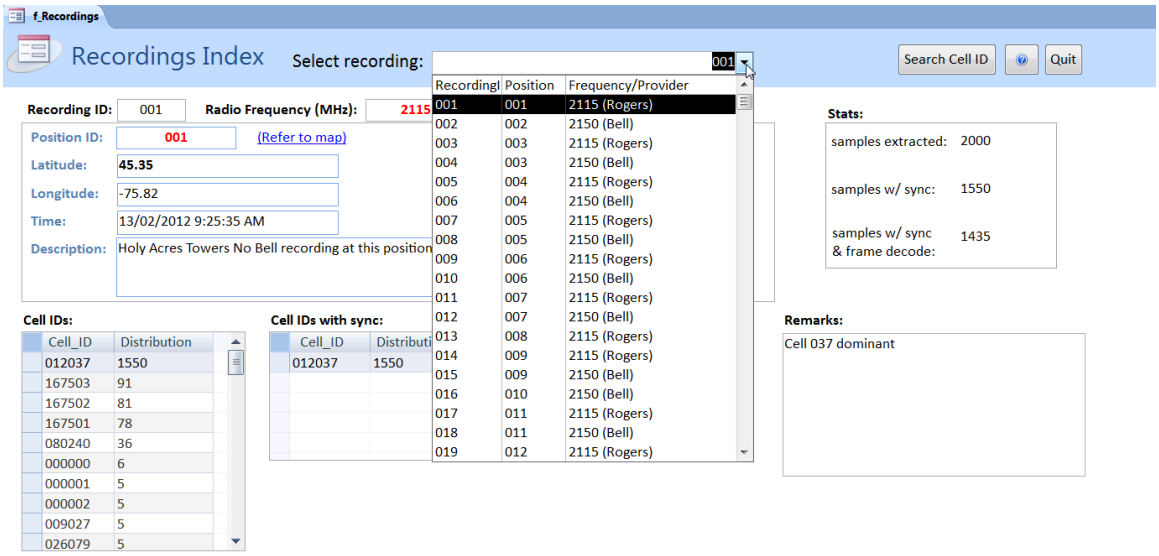



Figure 3.5: Screen capture of the *RecordingsIndex.accdb* database.

position markers in line with the Google Maps API. The output of this module was inserted into the HyperText Markup Language (HTML) code from [45] to overlay basic statistics of the recordings when the mouse hovers over a recording position. An example of position 25 is shown in Figure 3.6. In the example, two recordings were captured from position 25 (one per provider). Basic information about each recording is listed, such as the number of vectors with radio synchronization, and the number of vectors with both radio synchronization and a decoded frame number. The most frequently cell identifiers captured in each recording are also listed in order to view the reach of a particular cell across multiple recordings. The complete set of positions is available online at http://fdemers.com/LTE [46].

### 3.2.3 *analyze.sh*

This utility was purpose-built to perform the SVM analysis using the trace extracts as input, and constitutes the core of the analysis process depicted in Figure 3.2. It is used in conjunction with a large number of options configurable from the command line, as arguments. Notably, the utility was programmed to allow a large number of traces to be processed in a combined fashion, such that all trace extracts from one provider

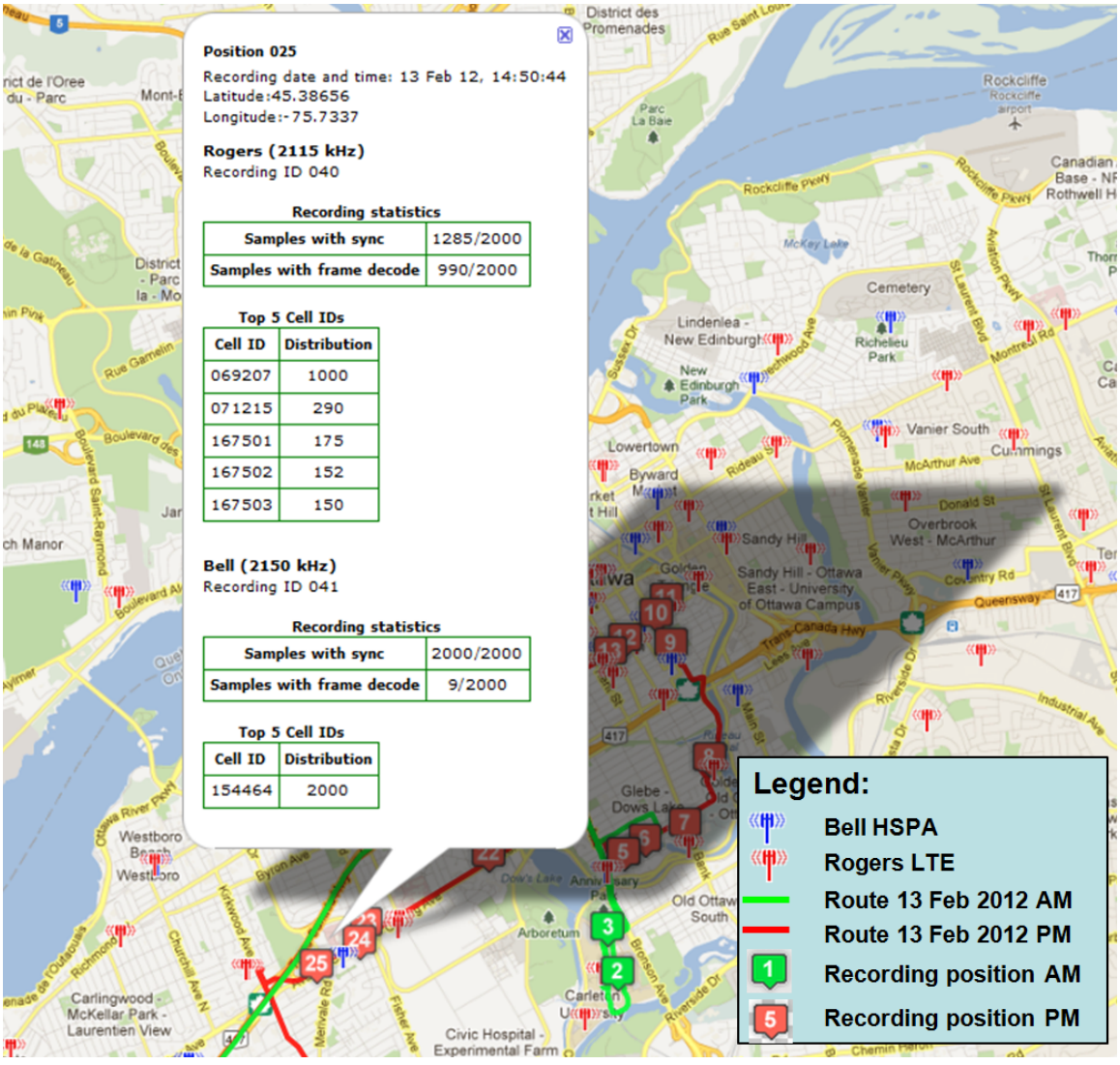



Figure 3.6: Geographical overlay of recording statistics.

could be considered as one. It further provides the ability to filter feature vectors that exhibited radio synchronization and/or a decoded LTE frame number. The command-line options programmed into *analyse.sh* are listed in Table 3.3. *analyze.sh* first processes the extracted feature vectors by removing unwanted columns from the trace, computing a new unique CellID as discussed in Section 4.1.3, and performs the filtering of vectors with synchronization and vectors with decoded frame number, as specified by the command line arguments. It then randomizes the vector set and reformats the trace according to the *libsvm* input format. *analyse.sh* then prunes cells with too few vectors (low density cells) and splits the population of vectors between a training set and a testing set. Scaling is the process used to normalize the numerical value range of each feature to ensure each feature variance is evenly considered in the SVM model. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation [35]. Both sets are scaled with the scaling factors obtained from the training set. When feature scoring is used, the features are weighted by the *weight.awk* script, based on either rank or score output by the chosen scoring algorithm (F-score or mRMR). Optionally, a grid search is performed to obtain optimal SVM parameters $C$ and $\gamma$ used with an RBF kernel function. Lastly, the core *svm-train* and *svm-predict* programs from [45] are invoked to obtain a re-identification success rate.

### 3.2.4 *looper.sh*

A simple Linux Bash script used to run a series of analysis tasks (by *analyze.sh*) based on the variables under study. This program took many forms throughout this research activity. The results were extracted from the *looper.sh* log file, a sample of which is shown in Figure 3.7, using *awk*, and entered into a Microsoft Excel spreadsheet from which tables and graphs were generated.

### 3.2.5 *fscore.py*

*fscore.py* is a Python implementation of the F-score feature scoring algorithm, written by Chen [41]. The F-score feature scoring algorithm is discussed in Subsection 3.4.1. The script outputs three files: *.fscore* shows importance of features, *.select* gives the

Table 3.3: *analyze.sh* options.

| Switch | Option | Description |
|---|---|---|
| usage: ./analyze.sh [-n #] [-t #] [-T #] [-s] [-f] [-i] [-r] [-o] [-S M/F] [-w 1-5] [-v] [-k] [-d] file1 file2 ... | | |
| -n | number of training vectors | Sets the number of vectors needed for the SVM training set. Cells with fewer vectors are normally excluded from the trial (default: 15). |
| -t | trials | Sets the number of trials to repeat the analysis several times (also sets -r) (default: 1). |
| -T | Trace | Sets the VSA trace number (default: 1 - Error Vector Summary trace) (other traces not implemented yet). |
| -s | sync | Keeps only vectors with SYNC and RAW SYNC flags set to TRUE (default: false). |
| -f | frame number decoded | Keeps only vectors with a decoded frame number (default: false). |
| -i | include | Includes low density cells the training model (default: false). |
| -r | randomize | Sorts the set of vectors in a random fashion before SVM analysis (default: false). |
| -o | optimize | Searches the best $C$ and $\gamma$ coefficients for the SVM analysis (default: false). |
| -S | scoring | Ranks and scores features using either mRMR (-SM) or F-score (-SF) algorithms (default: neither). |
| -w | weighting | Applies weights according to one of five algorithm (imples -SF if not specified, except for -w5). |
| -v | verbose | Outputs additional information concerning the progress of the analysis (default: false). |
| -k | keep | Keeps all temporary files (default: false). |
| -d | debug | Outputs additional information concerning the progress of the analysis (also sets -vk) (default: false). |

```
Cleaning up - removing temporary files
executing ./analyze.sh -v -r -SM -w2 -t5 -n10 ../data/002-DL-10MHz-2150MHz-10s-13Feb12-p002-n2000-o1300-d1-t1.csv ../data/004-DL-10MHz-2150MHz-10s-13Feb12-p003-n2000-o1300-d1-t1.csv
../data/006-DL-10MHz-2150MHz-5s-13Feb12-p004-n2000-o1300-d1-t1.csv ../data/008-DL-10MHz-2150MHz-5s-13Feb12-p005-n2000-o1300-d1-t1.csv ../data/012-DL-10MHz-2150MHz-5s-13Feb12-p007-n2000-o1300-d1-t1.csv
../data/015-DL-10MHz-2150MHz-5s-13Feb12-p009-n2000-o1300-d1-t1.csv ../data/016-DL-10MHz-2150MHz-5s-13Feb12-p010-n2000-o1300-d1-t1.csv ../data/018-DL-10MHz-2150MHz-5s-13Feb12-p011-n2000-o1300-d1-t1.csv
../data/020-DL-10MHz-2150MHz-5s-13Feb12-p012-n2000-o1300-d1-t1.csv ../data/023-DL-10MHz-2150MHz-5s-13Feb12-p014-n2000-o1300-d1-t1.csv ../data/025-DL-10MHz-2150MHz-5s-13Feb12-p015-n2000-o1300-d1-t1.csv
../data/026-DL-10MHz-2150MHz-5s-13Feb12-p016-n2000-o1300-d1-t1.csv ../data/028-DL-10MHz-2150MHz-5s-13Feb12-p017-n2000-o1300-d1-t1.csv ../data/029-DL-10MHz-2150MHz-5s-13Feb12-p018-n2000-o1300-d1-t1.csv
../data/032-DL-10MHz-2150MHz-5s-13Feb12-p020-n2000-o1300-d1-t1.csv ../data/035-DL-10MHz-2150MHz-5s-13Feb12-p022-n2000-o1300-d1-t1.csv ../data/037-DL-10MHz-2150MHz-5s-13Feb12-p023-n2000-o1300-d1-t1.csv
../data/039-DL-10MHz-2150MHz-5s-13Feb12-p024-n2000-o1300-d1-t1.csv ../data/041-DL-10MHz-2150MHz-5s-13Feb12-p025-n2000-o1300-d1-t1.csv ../data/043-DL-10MHz-2150MHz-5s-13Feb12-p026-n2000-o1300-d1-t1.csv
../data/045-DL-10MHz-2150MHz-5s-13Feb12-p027-n2000-o1300-d1-t1.csv

Processing 21 input files...

A total of 42000 vectors are combined for analysis (merged_inputs.csv)
Out of 171 unique cells, a population of 30 cells meets the requirement of having more than 10 vectors
(A total of 300 training vectors and 41256 testing vectors will be generated)
Starting trial #1
Randomizing vector set and converting trace format for SVM input (merged_inputs.svm)...
Populating training (training.svm) and testing (testing.svm) vector sets
Starting SVM analysis
 Scaling training vectors (generating scales.dat)
Scaling testing vectors (using scales.dat)
Using mRMR routine to rank features
 Normalizing class names for mRMR input (mrmr-input.csv)
 Invoking mRMR to generate feature scoring (feature-weights.csv)
 results of the mRMR scoring
-----------

6 1.111 RSRP
13 0.531 IQOffset
4 0.512 RSTP
5 0.506 OSTP
9 0.491 SyncCorr
1 0.4 EVM
7 0.362 RSRQ
3 0.297 RSEVM
8 0.262 FreqErr
2 0.257 EVMPeak
10 0.149 CTE
11 0.081 SymClkErr
12 0.002 TimeOffset

-----------
 Adjusting feature weights (algorithm=2)
Generating SVM model file (model.dat)
Computing SVM prediction of testing set (results.dat)
Accuracy = 76.6458% (31621/41256) (classification)
```

Figure 3.7: Sample output of the *analyze.sh* processing script.

running log, and *.pred* gives testing result [41].

### 3.2.6 *mrmr*

Compiled version of the minimum-redundancy-maximum-relevance (mRMR) feature selection algorithm, provided by Peng et al. [44]. As the *mrmr* program requires a different trace format, *analyze.sh* first proceeds by reformatting the data. A discussion on mRMR is available in Subsection 3.4.2.

### 3.2.7 *grid.py*

The *grid.py* Python script is provided by the *libsvm* library as a tool to find the optimal SVM parameters $C$ and $\gamma$ used with the RBF kernel function. It uses k-fold cross-validation and performs a grid search to determine the values of $C$ and $\gamma$ which result in the greatest prediction accuracy. Its usage is encouraged and explained by Hsu et al. in [35]. A discussion on the requirement for this search appears in Section 3.6.

### 3.2.8 *weight.awk*

*weight.awk* is a simple *awk* script used to multiply the values of each feature using one of five weighting algorithms. It uses, as input, the score or rank from either feature scoring algorithm. Weighting algorithms are discussed in detail in Section 3.5. The output of the two scoring algorithms had to be standardized in order to be compatible with the weighting algorithms.

### 3.2.9 Computing Cell IDs

The statistics regarding the data collected during the Ottawa city field trial are compiled in Table 4.1. The cell ID was computed using a concatenation of the CellId-GroupSector and the CellId fields from the trace extracts. Both of these entities are transmitted using a Zadoff-Chu sequence, in accordance with the LTE protocol. It was discovered that some cell IDs are used by both providers (both the CellId-GroupSector and CellId are identical). This explains why the number of cell IDs when combining all recordings ($3^{\text{rd}}$ column) is less than the total number of cell IDs if treating the two providers independently (last column) in Table 4.1.

## 3.3 Signal Quality Filtering

A signal quality filter was programmed into the analysis tool in order to allow filtering of the vector set based on two binary signal quality parameters, namely signal synchronization and decoded LTE frame number. Signal quality filtering was found to greatly improve emitter re-identification success rates, while at the same time significantly reducing the pool of suitable parameter vectors for the SVM analysis (see Table 4.1). Using command-line arguments during trace processing, it was possible to exclude vectors that had one or two of the parameters unset.

### 3.3.1 Signal Synchronization

The program used to step through the RF recording and extract parameter vectors (*VSA_Interface.xlsm*) also retrieved and tabled two distinct values labelled as signal synchronization using the documented Agilent API. The *AgtVsaVector.Application.Measurement.Status* was first recovered using the *vsaStatusBitSyncNotFound* bitmask (512) for every vector extracted, as well as the of the *AgtVsaVector.Application.Display.Traces(4).RawDataStatus* (where 4 corresponds to the trace number for the *D. Ch1 Error Summary* Table) using the *vsaTrcDatStaNoSync* bitmask (128). In practice, the data extracted showed that both signal synchronizations were equal, and thus redundant, for all vectors extracted. Interestingly, Agilent Technologies states in [42] that error measurements obtained in the absence of signal synchronization are not considered for the averaging, when averaging is enabled in the VSA software LTE demodulation properties. As such, it is recommended that at the very least, only feature vectors extracted in the presence of signal synchronization be considered for further study. Chapter 4 results confirm that re-identification results are severely impacted when feature vectors extracted in the absence of signal synchronization enter the SVM model, or are used as vectors under test.

### 3.3.2 Decoded Frame Number

In spite of the trace chosen for feature extraction, *VSA_Interface.xlsm* also captured the decoded downlink LTE frame number from the *Ch1 DL Decode Info* trace. Empirically, the decoded frame number was found to be a good indicator of received signal quality, as recordings with marginal signal quality prevented the VSA software from properly decoding the LTE downlink frame number.

## 3.4 Feature Ranking and Scoring

When presented a large number of features, it is reasonable to assume not all features are as useful in identifying unknown classes. Some features may present little variation from class to class, whilst others may present little information that has not already been characterized by another feature. A common strategy in classification problems is to make a determination which of the feature set should be emphasized, or even which feature set could be safely ignored. In order to do so, one must first determine "which" features, and next determine how to conduct this emphasis. The approach selected for this research activity consisted in trialing two feature scoring algorithms often used in classification problems, which attempt to answer the first question: F-score and mRMR. Secondly, feature weighting was chosen as a method by which emphasis could be applied, meaning most important features would be given a greater weight prior to being modelled by the SVM classifier, after normalization. Weighting can be seen as a more general case of feature exclusion, in which a set of $n$ top features is kept whereas the others are discarded. In feature weighting, this is the equivalent of setting unselected features to a constant, nullifying their contribution to the SVM classification model.

### 3.4.1 F-score

F-score is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors $x_k$, $k = 1, ..., m$, if the number of positive and negative instances are $n_+$ and $n_-$, respectively, then the F-score of the $i^{\text{th}}$ feature is defined as:

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{-} - \bar{x}_i^{(-)}\right)^2} \tag{3.2}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the average of the feature of the whole, positive, and negative data sets, respectively; $x_{k,i}^{(+)}$ is the $i^{\text{th}}$ feature of the $k^{\text{th}}$ positive instance, and $x_{k,i}^{(-)}$ is the $i^{\text{th}}$ feature of the $k^{\text{th}}$ negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative. A disadvantage of F-score is that is does not reveal

mutual information among features [47].

### 3.4.2 mRMR

One of the most popular approaches to realize Max-Dependency is maximal relevance (Max-Relevance) feature selection: selecting the features with the highest relevance to the target class $c$. Relevance is usually characterized in terms of correlation or mutual information, of which the latter is one of the widely used measures to define dependency of variables [44]. In feature selection, it has been recognized that the combinations of individually good features do not necessarily lead to good classification performance. In other words, "the $m$ best features are not the best $m$ features" [44]. Some researchers have studied indirect or direct means to reduce the redundancy among features and select features with the minimal redundancy (Min-Redundancy). The mRMR framework aims to minimize redundancy, used a series of intuitive measures of relevance and redundancy to select promising features for both continuous and discrete data sets [44].

Max-Dependency has the following form:

$$\max D\left(S,c\right), \quad D = I\left(\left\{x_i, i = 1, \cdots, M\right\}; c\right) \tag{3.3}$$

given the input data $D$ tabled as $N$ samples, and $M$ features $X = \{x_i, i = 1, \cdots, M\}$, and where $I$ is the mutual information between variables, for a given class $c$. As Max-Dependency criterion is hard to implement, an alternative is to select features based on maximal relevance criterion (Max-Relevance). Max-Relevance is to search features satisfying (3.4), which approximates $D\left(S,c\right)$ with the mean value of all mutual information values between individual feature $x_i$ and class $c$, for a feature set $S$ [44].

$$\max D\left(S,c\right), \quad D = \frac{1}{|S|}\sum I\left(x_i; c\right) \tag{3.4}$$

## 3.5 Feature Weighting and Selection

Once the features are ranked and scored by the scoring algorithm (either F-score or mRMR), it is possible to weight best features more heavily in an attempt to improve SVM prediction accuracy. A number of algorithms are possible using either the feature rank or feature score as input. The simplest approach is to multiply

linearly each vector element by the corresponding feature score. It is also possible to using the feature rank as a basis for the linear multiplication. Additionally, simple exponential functions using the score or the rank, which further emphasizes highly-ranked features, were also investigated.

A total of five weighting algorithms were implemented and used in this research activity, and presented in Table 3.4. For a given feature vector $\mathbf{a} = (a_1, a_2, \cdots, a_{13})$, the weighted feature vector is denoted $\mathbf{w} = (w(a_1)\, a_1, w(a_2)\, a_2, \cdots, w(a_{13})\, a_{13})$. In the table below, the function $f(a_i)$ returns the score obtained by feature $a_i$ from the scoring algorithm (F-score or mRMR) whereas the function $F(a_i)$ returns a value proportional to $R(a_i)$, the rank of parameter $a_i$ ($F(a_i) = 13$ if $a_i$ is ranked first, $F(a_i) = 1$ if $a_i$ is ranked last). Chapter 4 studied all possible combinations of feature scoring and feature weighting algorithms to determine which one performs best.

Table 3.4: Weighting algorithms.

| Algorithm number | Description | Formula |
|---|---|---|
| 1 | Multiplying each feature by the corresponding value obtained from the scoring algorithm. | $w(a_i) = f(a_i)$ |
| 2 | Multiplying each feature by the corresponding rank obtained from the scoring algorithm. | $w(a_i) = F(a_i)$ |
| 3 | Multiplying each feature by 2 to the power of the corresponding value obtained from the scoring algorithm. | $w(a_i) = 2^{f(a_i)}$ |
| 4 | Multiplying each feature by 2 to the power of the corresponding rank obtained from the scoring algorithm. | $w(a_i) = 2^{F(a_i)}$ |
| 5 | Brik's special case: all features ignored except error vector magnitude ($a_1$), I&Q origin offset ($a_{13}$), frequency error ($a_8$), and SYNC correlation ($a_9$). | $w(a_i) = \begin{cases} 1, \forall i \in \{1, 8, 9, 13\} \\ 0, \quad \text{others} \end{cases}$ |

We treat feature selection as a special case of feature weighting in which selected

features are weighted with unity (i.e. retained) and all others ignored (zeroed). The most useful application of feature selection is to retain only $n$ features that are the most effective at re-identification, after scoring as in Equation (3.5). As an example, and in order to compare our results with the results obtained by Brik et al. in [2], we attempted to compute re-identification accuracy using the same five parameters, namely (from the most effective to the least effective) (i) frequency error, (ii) SYNC correlation, (iii) I&Q offset, (iv) magnitude error, and (v) phase error. However, the phase error was not available in the VSA trace used. Thus the Brik comparison experiments ignored all but four parameters, using the feature selection scheme outlined in Table 3.4. Note that feature scoring was not required in the experiments using the fifth weighting algorithm, as it does not depend on either the score or the rank.

$$w\left(a_i\right) = \begin{cases} 1, & \forall i \in R\left(a_i\right) \leq n \\ 0, & \text{others} \end{cases} \tag{3.5}$$

## 3.6 SVM Parameters

Support vector machines attempt to find the solution to the optimization problem shown below

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to:} \quad & y_i\left(\mathbf{w}^T\phi\left(\mathbf{x}_i\right) + b\right) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{3.6}$$

where $C > 0$ is the penalty parameter of the error term, also termed regularization parameter. The kernel function $K$ is defined in Equation (3.7), and is the function used to express the hyperplane between classes. SVM, and *libsvm* in particular, offer a number of choice of kernels functions: linear, polynomial of $n$ degrees, RBF and sigmoid. The RBF kernel function is shown in Equation (3.8), where $\gamma$ is the kernel parameter.

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \tag{3.7}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \gamma > 0 \tag{3.8}$$

There are two adjustable parameters when using an RBF kernel: $C$ and $\gamma$. It is not known beforehand which $C$ and $\gamma$ are best for a given problem; consequently some kind of model selection (parameter search) must be done. The goal is to identify good parameters $C$ and $\gamma$ so that the classifier can accurately predict unknown data (i.e. testing data) [35]. The grid search method is recommended by Hsu in [35]. Various pairs of $C$ and $\gamma$ values are tried and the one with the best cross-validation accuracy is picked. Hsu found that trying exponentially growing sequences of $C$ and $\gamma$ is a practical method to identify good parameters. In order to reduce the computational requirements, a coarse grid search first identifies a suitable regions, which is followed by a finer grid search within the suitable regions. *libsvm* provides a grid search implementation that can be used to find optimal $C$ and $\gamma$ for a given training set.

### 3.6.1 *k*-fold Cross-Validation

Prediction accuracy is often used as the optimization criteria for the $C$ and $\gamma$ kernel parameters. At its basis, the training dataset is randomly separated into two parts, of which one is considered unknown. The prediction accuracy obtained from the "unknown" set against the second part reflects the likely performance on classifying an independent dataset. An improved version of this procedure is known as $k$-fold cross-validation. In $k$-fold cross-validation, the training dataset is first randomly divided into $k$ subsets of equal size. Sequentially, one subset is tested using the classifier trained on the remaining $k-1$ subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which is correctly classified. The cross-validation procedure can prevent the overfitting problem [35].

## 3.7 Performance Evaluation

The SVM prediction accuracy, output by *svm-predict* was considered to be a relevant measure of re-identification accuracy. *svm-predict* is able to compare the predicted

class name with the actual class name which is also present in the dataset. It is also possible to conduct in-depth analysis looking into which cells are most often misidentified by the SVM model, or conversely which cells are most often incorrectly selected by the SVM model. This knowledge could highlight the need to further discriminate between cells that have similar radiometric identities, using other methods. As this in-depth analysis is only relevant to a specific data set and does not apply well to other experiments investigating different transmitters, it was deemed out of scope. Additional information concerning performance evaluation criteria is available in [2].

# Chapter 4

# Results and Analysis

This chapter details the data collection and analysis process, as well as the results obtained during the course of this research activity. In the following section, details of the data collection, both from the lab environment and during the field trials, are presented. Results of the trace analysis are presented next. Attention was devoted to examine the impact of filtering, feature scoring and weighting, as well as SVM parameters optimization on prediction accuracy, in line with what was presented in Chapter 3. An analysis of the significance of the results concludes this chapter.

## 4.1 Experimental Procedures

### 4.1.1 Vector Signal Analysis

The Agilent N9030A PXA Signal Analyzer is a modern local area network (LAN)-enabled spectrum analyzer with capabilities in the 3 Hz to 13.6 GHz frequency range. It is compatible with the Agilent (89600B) VSA software. The current version of the Agilent VSA software package (89600B) includes both the VSA software (89601B) and the wireless link analysis (WLA) software (89620B), the MAC-layer complement to the 89601B. The WLA was not required for this research activity. The LTE frequency division duplexing (FDD) modulation Analysis (89601B-BHD) option was required, and fortunately was part of the software suite available at Defence Research and Development Canada (DRDC) Ottawa. This optional package provides the necessary algorithms and protocol knowledge to analyze LTE FDD transmissions. Figure 4.1 shows a sample workspace of the Agilent VSA software during an experiment, processing an LTE FDD downlink signal in the City of Ottawa in December, 2011. The

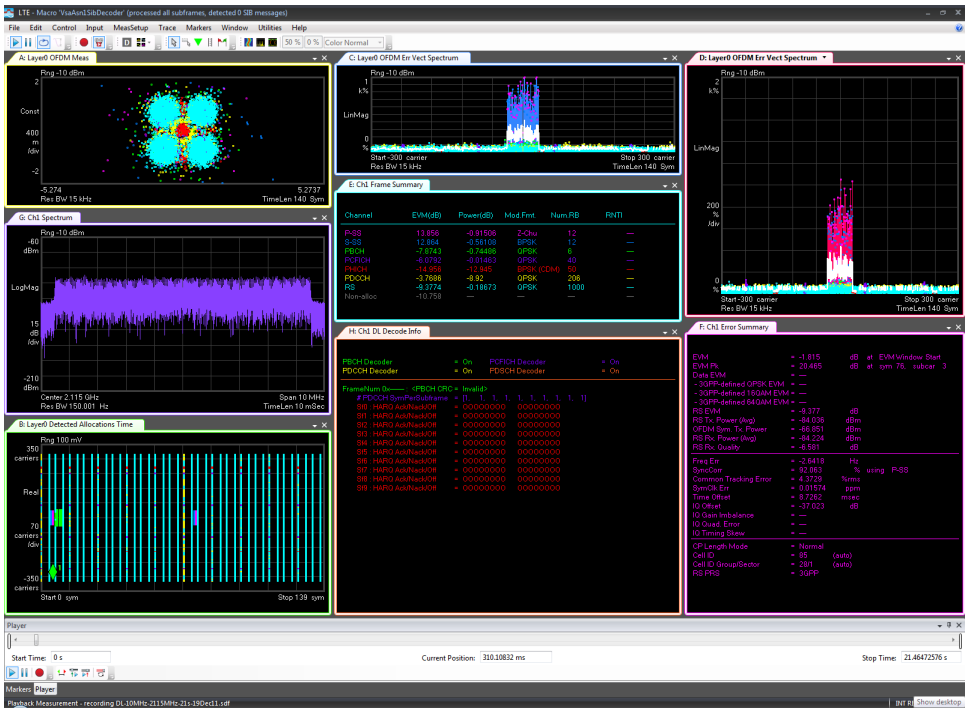


Figure 4.1: Screen capture of the Agilent VSA software processing an LTE signal in the City of Ottawa.

figure also shows the *Error Summary* table (bottom right) which provided the values extracted into feature vectors for the SVM classifier.

### 4.1.2 DRDC Ottawa Laboratory

Early into the research activity, the decision was made to make use of the test and measurement instrumentation at the DRDC Ottawa campus at Shirley's Bay. The Modern Communications Electronic Warfare (MCEW) group agreed to make the Agilent N9030A PXA Signal Analyzer as well as the VSA software package available for this research activity. A number of antennas and ancillary equipment were also made available as needed.

Initially, a set of recordings were performed from the Shirley's Bay laboratory between November 2011 and February 2012. The recordings features were extracted and the hypotheses verified (it is possible to uniquely distinguish LTE emitters using radiometric properties) with a limited population of cellular towers. The laboratory experiment consisted of the Agilent N9030A PXA Signal Analyzer connected

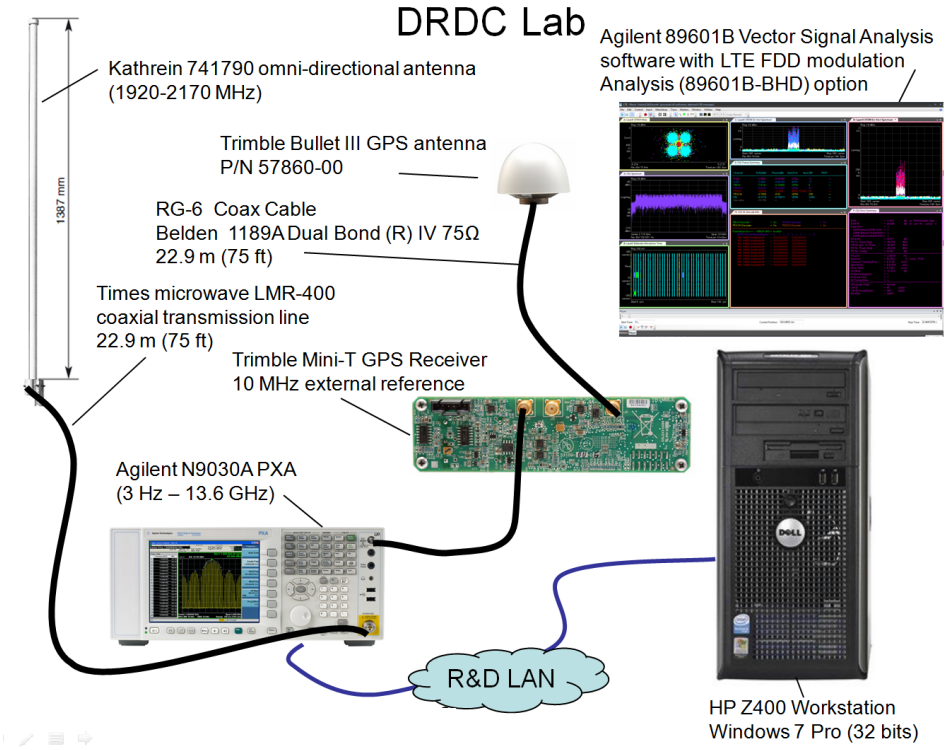



Figure 4.2: Laboratory assembly.

to Kathrein 741790 omni-directional antenna installed on the roof of the building, through 22.9 m of Times Microwave LMR-400 coaxial transmission line. The external 10 MHz reference was provided by a Trimble Mini-T GPS Receiver board fed from a roof-mounted Trimble Bullet III GPS antenna (P/N 57860-00) through 22.9 m of Belden 1189A Dual Bond (R) IV 75 Ω. An HP Z400 workstation with Windows 7 Professional (32 bits) ran the 32 bit version of the Agilent (89600B) VSA software (version 14.23). Figure 4.2 depicts the laboratory installation.

A total of 17 cells were observed from recordings at the Shirley's Bay, between the two service providers who had active FDD LTE signals (Bell Canada and Rogers). Re-identification success rate peaked, after feature scoring and SVM kernel optimization, at 91%. However, the data collected during this phase did not include information whether the LTE downlink frame number was properly decoded or the VSA's SYNC status. As discovered in subsequent experiments, having the ability to filter feature vectors based on signal quality improves the re-identification success rate.

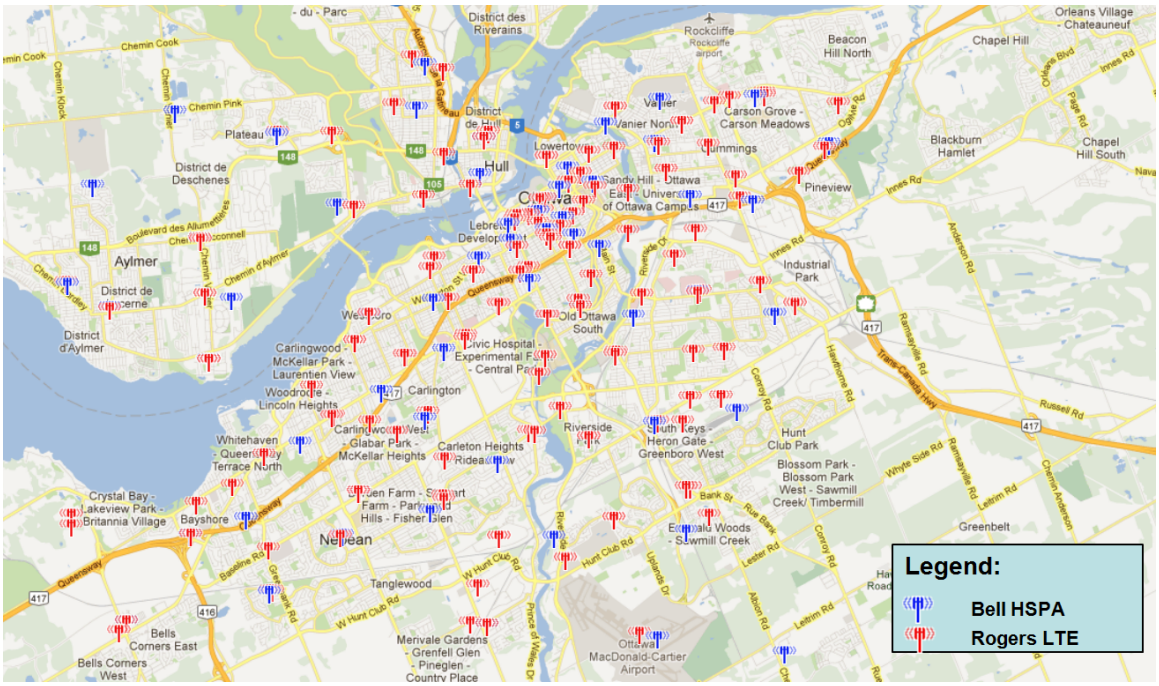



Figure 4.3: Cellular tower locations in Ottawa, with data from [48].

### 4.1.3 Recordings from the City of Ottawa

Given the small population of cells within radio range to the Shirley's Bay campus, the decision was made to perform recordings from various positions in the Ottawa area. A route was planned around the City of Ottawa in order to increase the number of candidate cells and determine if a larger population would harm re-identification. Route planning was conducted using cell tower positions available from [48]. Note that [48] listed specific Rogers LTE tower positions but did not specify Bell LTE tower positions. It was assumed, for the purpose of route planning, that Bell co-located LTE towers with the HSPA towers. Tower positions around Ottawa for the two providers are amalgamated in Figure 4.3. The visualization service GPSvisualizer.com [45] was used to generate the map, which was then modified using a text editor.

Since measurements would be taken only when the vehicle was at rest, the following factors were considered during route planning:

- Locations close to cell towers from two providers were favoured. This enabled multiple recordings from a single position. The proximity to the towers also increases the quality of the signal recording.

- A high concentration of towers were available in the down-town core. Thus a significant number of the stops were planned within the down-town area.
- In order to test the effect of signal-to-noise ratio (SNR) on re-identification success rate, the route was planned to record signals from certain towers multiple times, varying the distance to the towers.

Microsoft MapPoint with GPS dongle was used to confirm our position in relation to reported cell towers from [48]. A Garmin eTrex Vista HCx was used to record the geographical coordinates of each recording position. The Agilent N9030A PXA Signal Analyzer as well as the workstation with the VSA software package was removed from the laboratory and secured within a DRDC vehicle to perform the signal recordings. The PCTEL BLMPVDB700/2500 antenna with the BGMML195MSMA magnetic base was installed on top of the van and connected to the PXA using 3.6 m of Pro-Flex Plus 195 coaxial cable. Again, the external 10 MHz reference was provided by a Trimble Mini-T GPS Receiver board fed from a Trimble Bullet III GPS antenna P/N 57860-00 through 22.9 m of Belden 1189A Dual Bond (R) IV 75 Ω. The GPS antenna remained within the vehicle. A direct RJ-45 LAN connection was established between the workstation and the PXA. Attempting to run the equipment using the on-board alternator through an Absopulse sine-wave inverter resulted in abrupt power failures at low engine revolutions per minute (RPM). This issue forced the research team to stop and resume the experiment using a Honda 2 kW generator secured to a tray attached to the trailer-hitch. Figure 4.4 shows the experimental assembly for this portion.

The recordings were conducted on February 13$^{\text{th}}$, 2012. Both LTE cellular providers in the area were active by that date. A total of 36 recording positions generated 54 recordings, as both providers did not broadcast LTE signals from every position. The route followed, and each recording position, are overlaid on the tower positions in Figure 4.5. A summary of the number of feature vectors extracted for each provider is found in Table 4.1. For every recording, a set of 2000 feature vectors was extracted from the VSA software using *VSA_Interface.xlsm*. From the 2000 vectors, a fraction had radio sync (94.6%) and a smaller fraction had both radio sync and a properly decoded frame number (30.8%).

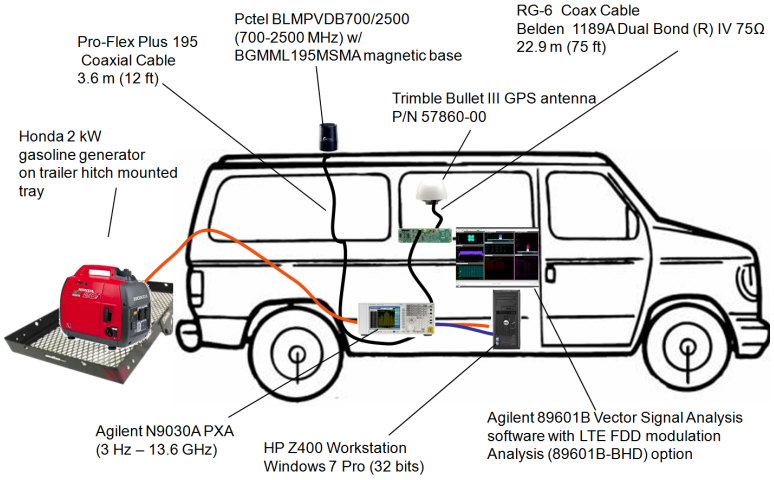



Figure 4.4: Mobile laboratory assembly.

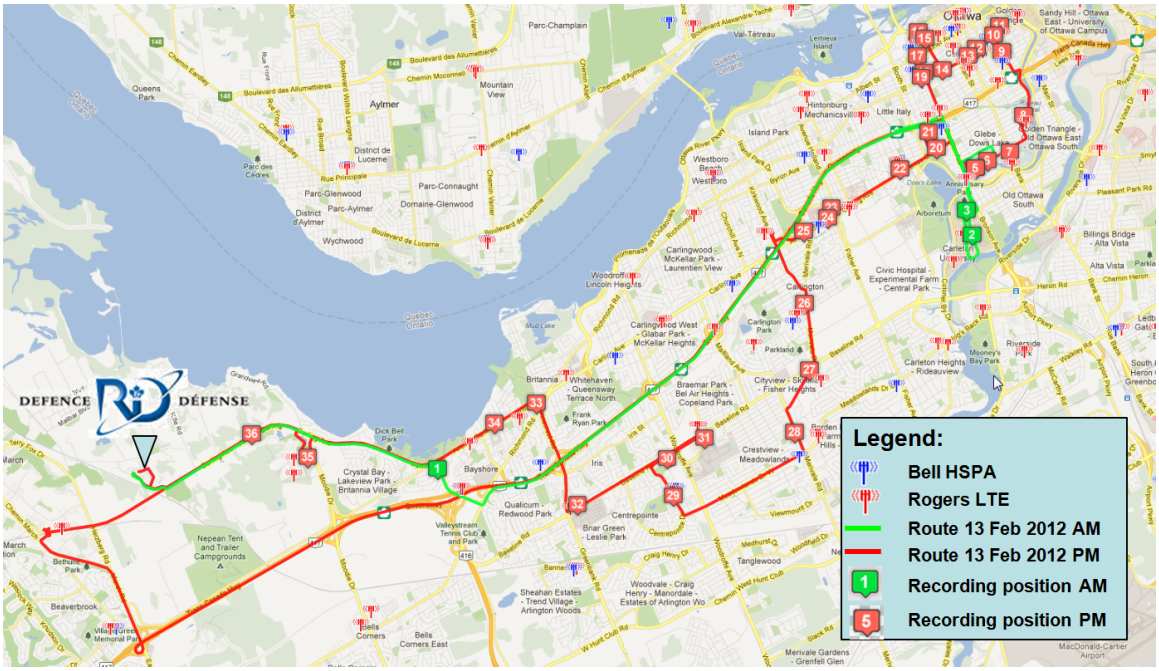



Figure 4.5: Cellular towers in Ottawa and data gathering route, with data from [48].

Table 4.1: Statistics from the recordings in the City of Ottawa 2012.

| **Filter** | **Elements** | **Bell** | **Rogers** | **Combined** | **Total** |
|---|---|---|---|---|---|
| | Recordings | 22 | 32 | 54 | 54 |
| no filtering | Vectors collected | 44000 | 64000 | 108000 | 108000 |
| | Unique Cell IDs | 173 | 346 | 396 | 519 |
| with radio sync | Vectors collected | 42214 | 59960 | 102174 | 102174 |
| | Unique Cell IDs | 83 | 203 | 250 | 286 |
| with frame # decoded | Vectors collected | 18923 | 14302 | 33225 | 33225 |
| | Unique Cell IDs | 18 | 39 | 55 | 57 |

## 4.2 Experimental Results

This section presents experimental results obtained and discusses their impact on practical applications and future research. For every experiment, the dataset obtained from the VSA software was divided into a small training set and a much larger testing set. *svm-train* generates a model file from the training set. *svm-predict* attempts to determine the class (or cell ID) of each vector in the testing set.

The dataset used as input consisted of the entire set of vectors for a given cellular provider. In general, traces from both providers were not combined because of duplicate cell IDs (see Subsection 3.2.9), except for work in Subsection 4.2.11 which specifically examines the impact of merging feature vector sets from both providers. For experiments using a random sorting of feature vectors prior to splitting the training and testing sets, experimental results consist of the prediction accuracy averaged over 5 experiments, with graphs showing error bars representing the 95% confidence interval. For the experiments which did not first randomize the trace data, the results consist of the prediction accuracy over a single experiment since repeat experiments returned the same results.

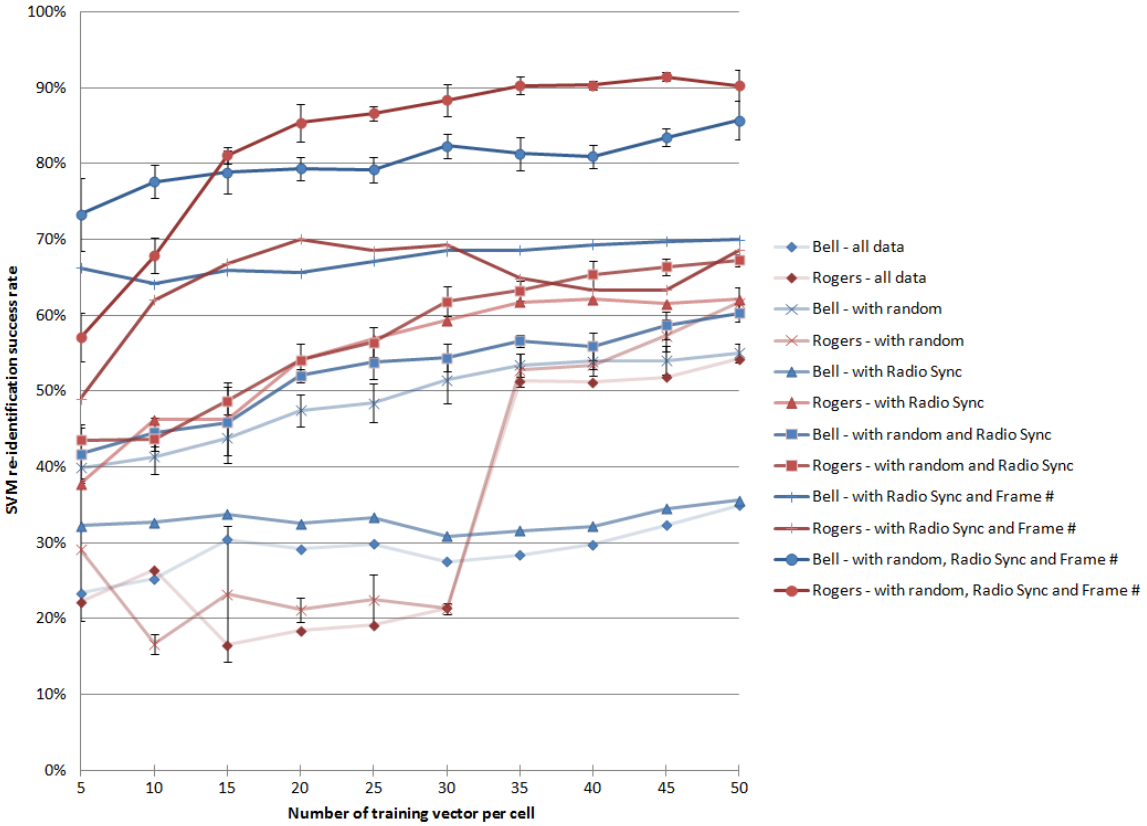



Figure 4.6: Prediction accuracy as a function of the number of training vectors, randomness and signal filtering.

### 4.2.1 Signal Quality Filtering

Filtering of the dataset based on signal quality parameters (radio synchronization and decoded LTE frame number) were found to have a tremendous impact on prediction accuracy, consistent with expectations. It is believed that higher signal quality leads to improved accuracy of the SVM model and in turn improves re-identification success rates. Previous research activities, notably Brik et al. in [2] operated from a controlled classroom environment rather than outdoor field trials using operational commercial infrastructure and as such may not have had to concern their experiment with signal quality. Figure 4.6 shows the tremendous improvement in prediction accuracy when signal quality filtering, against one or two criteria, is applied, for both cellular providers. The re-identification accuracy peaks at 55% and 61% (with randomness) without filtering, whereas success rates of 91% and 85% are obtained under the strictest filtering.

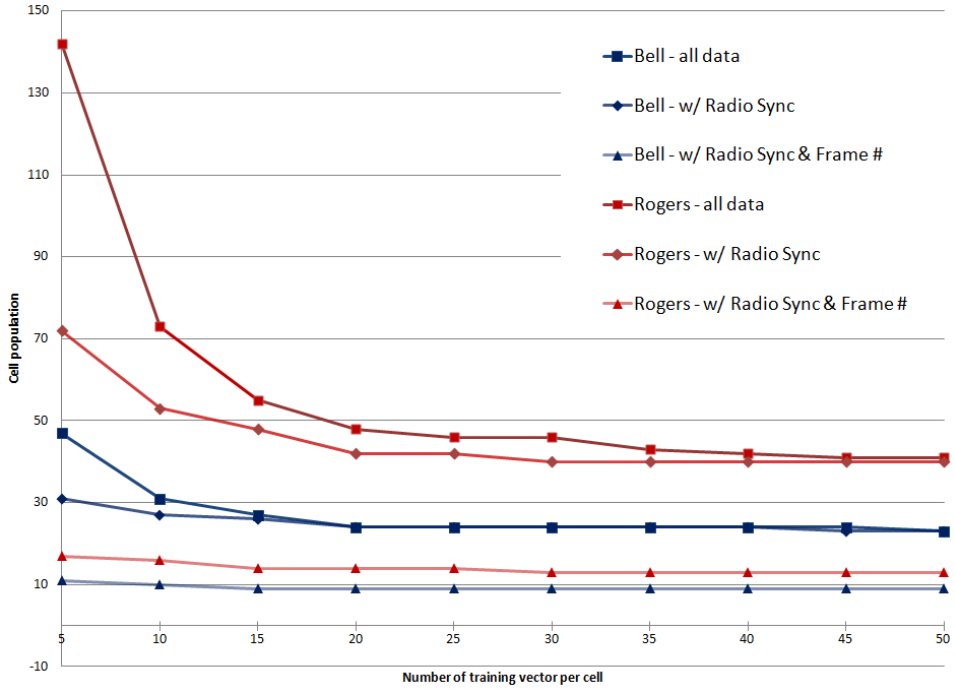



Figure 4.7: Cell population as a function of the training bin size.

### 4.2.2 Variation of the Training Bin Size

The number of vectors (per unique cell) used in the training set was found to have a significant impact on re-identification accuracy, in line with expectations. The training bin size was varied from 5 vectors to 50 vectors (in increment of 5), per unique cell. Figure 4.6 also shows the improvements in prediction accuracy for larger training sets. The re-identification accuracy peaks at 73% and 57% for a set of 5 training vectors, whereas success rates of 91% and 85% are obtained when the set is 45 and 50 training vectors, respectively. The larger training bin size also impacted upon the number of candidate cells retained, particularly if filtering on radio synchronization and decoded frame number were enabled. This is because fewer cells had a sufficiently high number of vectors to be retained in the experiment. Table 4.2 shows the number of candidate cells retained as a function of the number of training vectors selected and the trace filtering applied, based on signal quality. Figure 4.7 shows the same information graphically, which demonstrates that in spite of cell population stabilizing, greater re-identification accuracy is still achieved when a greater number of training vector is considered to build the SVM model.

Table 4.2: Number of candidate cells as a function of filtering and number of training vectors.

| | # cells retained | | | | | |
|---|---|---|---|---|---|---|
| | No filtering | | Radio sync | | Radio sync & frame # decoded | |
| # training vectors | Bell | Rogers | Bell | Rogers | Bell | Rogers |
| 5 | 47/173 | 142/346 | 31/83 | 72/203 | 11/18 | 17/39 |
| 10 | 31/173 | 73/346 | 27/83 | 52/203 | 10/18 | 16/39 |
| 15 | 27/173 | 55/346 | 26/83 | 48/203 | 9/18 | 14/39 |
| 20 | 24/173 | 48/345 | 24/83 | 42/203 | 9/18 | 14/39 |
| 25 | 24/173 | 46/346 | 24/83 | 42/203 | 9/18 | 14/39 |
| 30 | 24/173 | 46/346 | 24/83 | 40/203 | 9/18 | 13/39 |
| 35 | 24/173 | 43/346 | 24/83 | 40/203 | 9/18 | 13/39 |
| 40 | 24/173 | 42/346 | 24/83 | 40/203 | 9/18 | 13/39 |
| 45 | 24/173 | 41/346 | 23/83 | 40/203 | 9/18 | 13/39 |
| 50 | 23/173 | 41/346 | 23/83 | 40/203 | 9/18 | 13/39 |

It is believed that a higher number of vectors in the training bin improves the accuracy of the SVM model and in turn improves re-identification success rates. Brik et al. had found in [2] that in practice, 20 frames were sufficient for optimized re-identification. The data shown in Figure 4.6 shows that results continue to improve up to 50 training vectors (per cell). In some cases, a larger training could lead to model overfitting and decrease prediction accuracy, as appears to be the case for the Rogers dataset, with randomness and strict signal filter, between 45 and 50 training vectors. However, the error bars show that the decrease is not statistically significant. Higher number of training vectors were not attempted in order to avoid reducing the population of candidate cells to almost trivial levels.

### 4.2.3 Randomness

Randomness of the dataset was found to have a tremendous impact on re-identification accuracy. Randomizing the dataset is the recommended approach prior to any *libsvm*

analysis in the practical guide by Hsu et al. [35], as well as in the Waikato Environment for Knowledge Analysis (WEKA) tutorial [49]. Figure 4.6 shows the considerable improvements in prediction accuracy when the input dataset is first randomized. Without randomizing, the re-identification accuracy peaks at 70% and 69%, whereas success rates of 91% and 85% are obtained when the vector set is first randomized. Furthermore, a number of anomalies were observed in experiments where randomness was not used (see Figures 4.7, 4.9, 4.13 and 4.15). Without randomness, the model is constructed using the first $n$ feature vectors where the cell ID is observed (trace extracts are fed to the model in order of recording positions), which may not be sufficiently representative of the radiometric properties of the cell. It is suspected that the SVM model suffers from inaccuracies introduced by the recording position where cells are first observed, which makes the model less universal than when it is constructed using feature vectors that originate from a broader set of cells, randomly selected from all recordings of a given provider.

### 4.2.4 Optimization of SVM Parameters

Optimization of the SVM parameters $C$ and $\gamma$ by grid search, using *k-fold validation* as discussed in Section 3.6.1, was found to have a significant positive impact on re-identification accuracy, in line with expectations and recommended *libsvm* usage promulgated by Hsu et al. in [35]. When SVM parameter optimization is enabled, *libsvm* produces a grid search graph using *gnuplot* such as the one shown in Figure 4.8. The optimal values of $C$ and $\gamma$ are then used when building the SVM training model. Figure 4.9 shows the improvements in prediction accuracy when SVM parameters are optimized using grid search, peaking at 98.9% and 98.0% under strictest signal filtering conditions and randomness. The anomaly observed in Figure 4.9 can perhaps be explained by the lack of randomness, suspected to cause the SVM model to optimize for a specific set of feature vectors which ill-represent the radiometric identity of cells when all trace extracts are considered at random.

Figure 4.10 compares the optimal curves with and without SVM parameter optimization and is perhaps more indicative. The graph shows the increased prediction accuracy as well as the smaller confidence interval, when searching for and applying the optimal SVM parameters $C$ and $\gamma$.

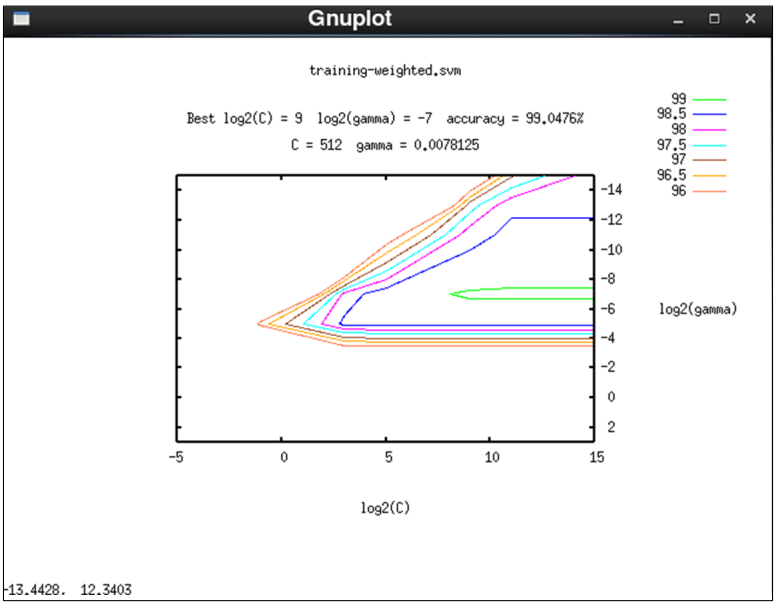



Figure 4.8: Optimization of SVM parameters $C$ and $\gamma$ by grid search.

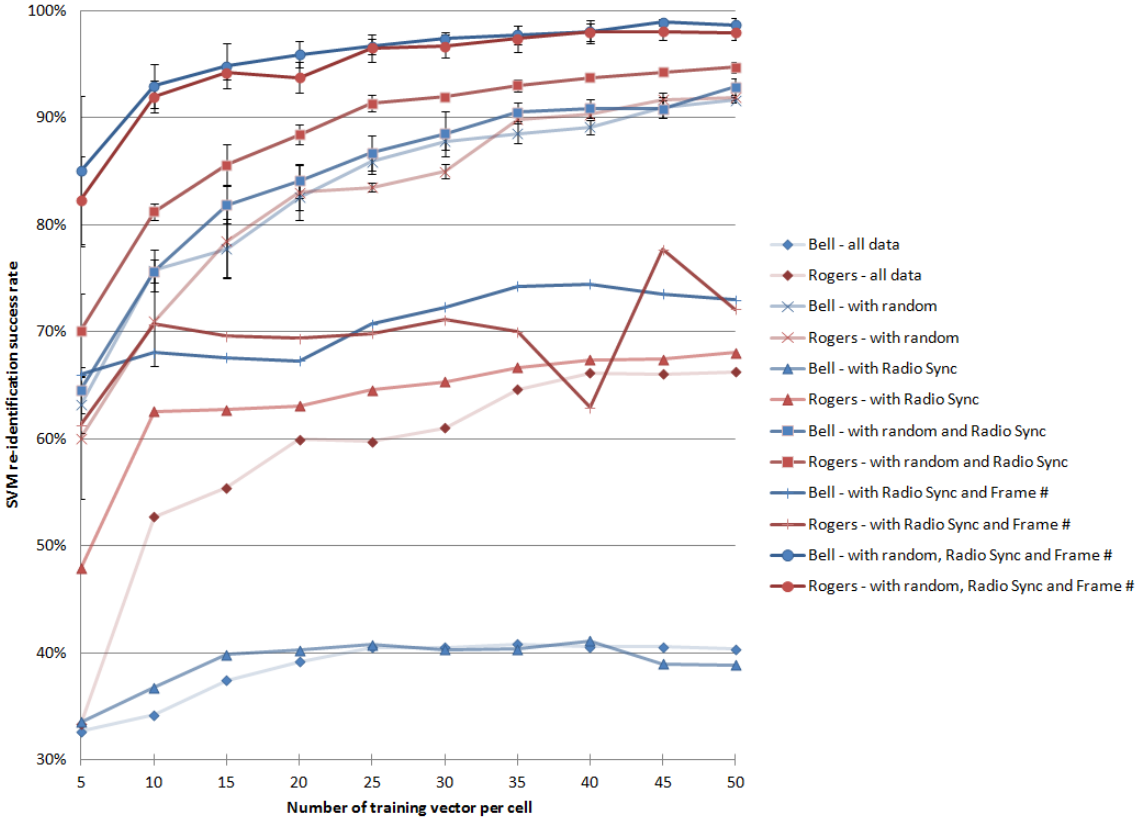



Figure 4.9: Prediction accuracy as a function of the number of training vectors, randomness and signal filtering, under SVM parameter optimization.

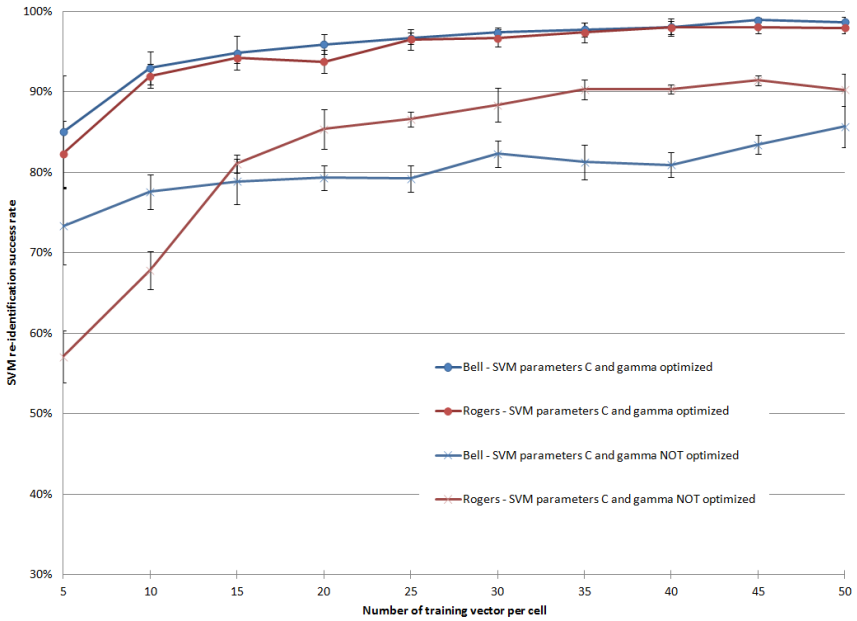



Figure 4.10: Prediction with and without SVM parameter optimization, under strictest signal filtering and randomness.

### 4.2.5 Inclusion of Low Density Cells

Under normal conditions, cells with fewer than the required minimum number of feature vectors are discarded. For example, when the training bin size is set to 45 vectors, all cells with fewer than 45 vectors are normally discarded from both the training and testing set. As an experiment, an option was added to *analyse.sh* to retain vectors belonging to low-density cells in the training set. Low density cells are defined as cells for which there are insufficient feature vectors to be included in the training model (the number of feature vectors for a given cell ID is less than $n$). In these experiments, feature vectors for low density cells are also included in the SVM model. The testing set never contained feature vectors for low density cells (all were used for the training set). Inclusion of low density cells in the SVM model were found to degrade the prediction accuracy, in line with expectations. It is suspected that the inclusion of these low-density cells increases the cell population and reduces the hyperspace between classes, rendering re-identification more difficult. Figure 4.11 shows the decrease in prediction accuracy when low density cells are retained for the model generation by the *svm-train* program. The results are sufficiently poor in some cases, to render impractical any re-identification applications, particularly if the low

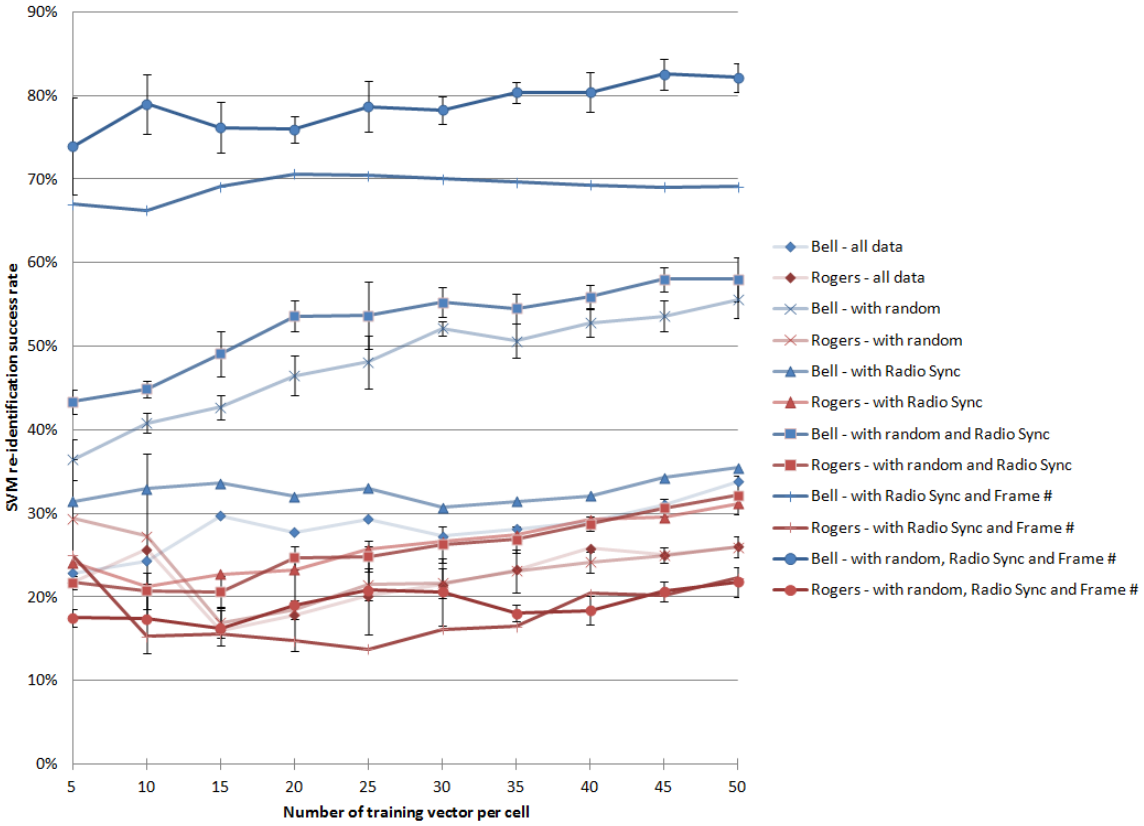



Figure 4.11: Prediction accuracy as a function of the number of training vectors, randomness and signal filtering, when including cells not characterized by the SVM model.

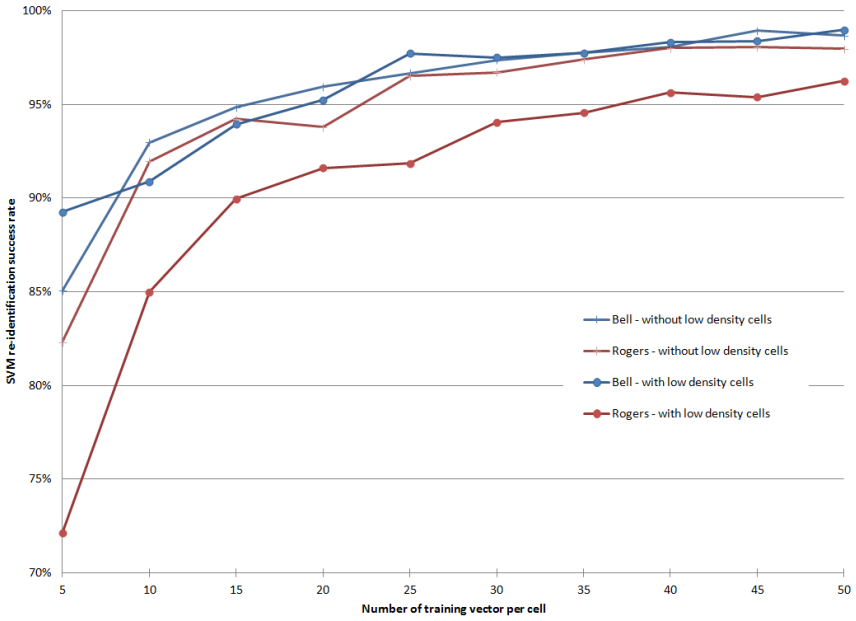



Figure 4.12: Prediction accuracy compared with and without low density cells, using SVM parameter optimization, randomness, presence of radio sync and a decoded frame number.

density cell population is large (as in the case of Rogers).

To facilitate the comparison with earlier results, Figure 4.12 shows re-identification results obtained with and without the inclusion of low density cells in the SVM model generation, with SVM parameter optimization for both providers. Error bars are not shown for clarity. The figure shows that SVM parameter optimization can effectively compensate some of the poor results obtained in the previous figure. It can be seen that the inclusion of low density cells reduce the re-identification accuracy unevenly for both providers. The variation for Bell is not statistically significant.

### 4.2.6 F-score Feature Scoring and Weighting

Results in this section are obtained using the feature scoring tool provided with *libsvm* by Chen and Lin in [41]. Feature scoring using the F-score algorithm resulted in the ranking found in Table 4.3. The ranking favoured features akin to the empirical results of Brik et al. in [2]. In this section, feature weighting was accomplished by multiplying each vector element by the corresponding feature score (weighting algorithm #1),

although *weight.awk* supported other weighting algorithms. The investigation of re-identification accuracy as a function of the $n$ best features, varying $n$ from 1 to 13, and excluding other features, has not been conducted and is recommended for future work.

Table 4.3: Sample feature scoring results using the F-score algorithm. See Section 3.1 for details.

| Feature number | Score | Feature description |
|---|---|---|
| 8 | 173.732404 | Frequency Error |
| 13 | 7.935522 | I &Q Offset |
| 4 | 5.775424 | RSTP |
| 6 | 4.974973 | RSRP |
| 3 | 1.733571 | RSEVM |
| 1 | 1.719871 | EVM |
| 2 | 1.361014 | EVM Peak |
| 5 | 0.884275 | OFDM symbol transmit power (OSTP) |
| 9 | 0.859362 | SYNC Correlation |
| 7 | 0.345288 | RSRQ |
| 12 | 0.107173 | Time Offset |
| 10 | 0.089607 | CTE |
| 11 | 0.020336 | Symbol Clock Error |

Figure 4.13 shows prediction accuracy when the F-score feature selection algorithm is used, and vector parameters are weighted linearly with the feature score output from the F-score algorithm, in line with expectations. In Figure 4.13, SVM parameters $C$ and $\gamma$ were optimized by grid search. The anomaly can perhaps be explained by the lack of randomness, suspected to cause the SVM model to optimize for a specific set of feature vectors which ill-represent the radiometric identity of cells when all trace extracts are considered at random. Again, it is shown that prediction accuracy improves under the most stringent signal quality filtering, using radio synchronization and decoded frame number, and using a large number of feature vectors during the

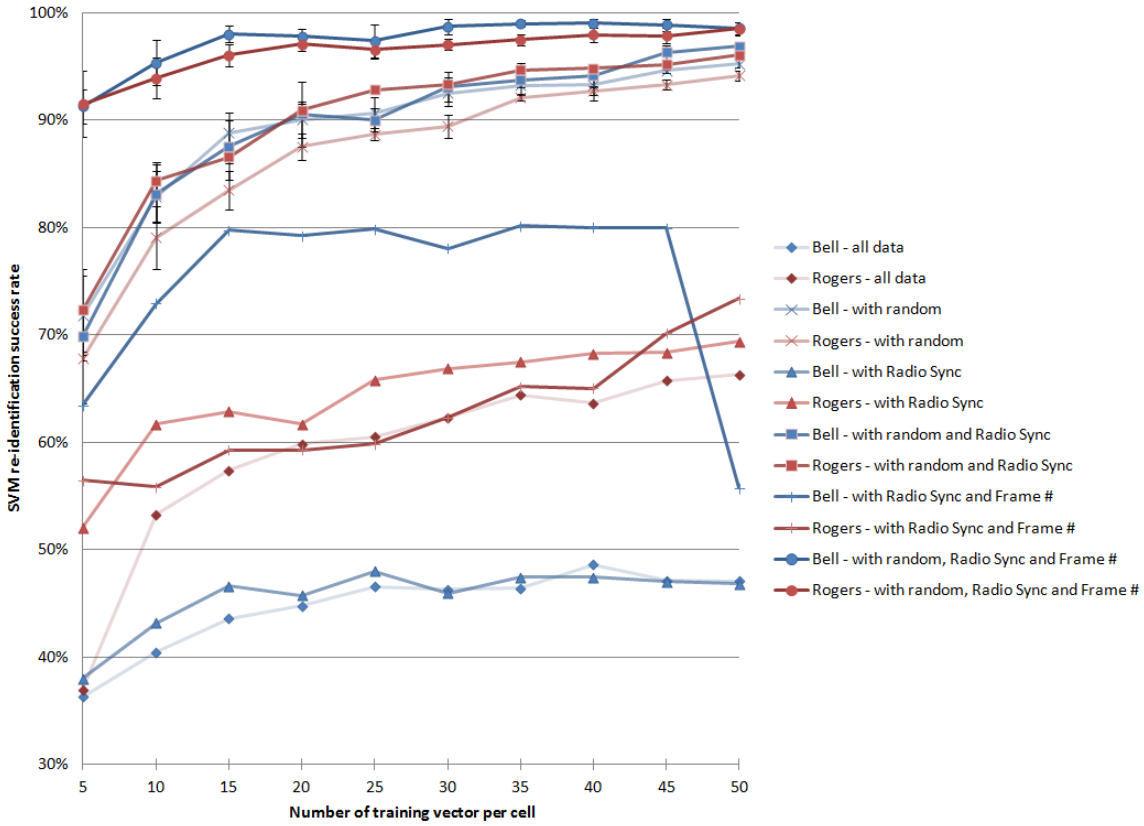



Figure 4.13: Prediction accuracy as a function of the number of training vectors, randomness and signal filtering, with F-score feature scoring and weighting.

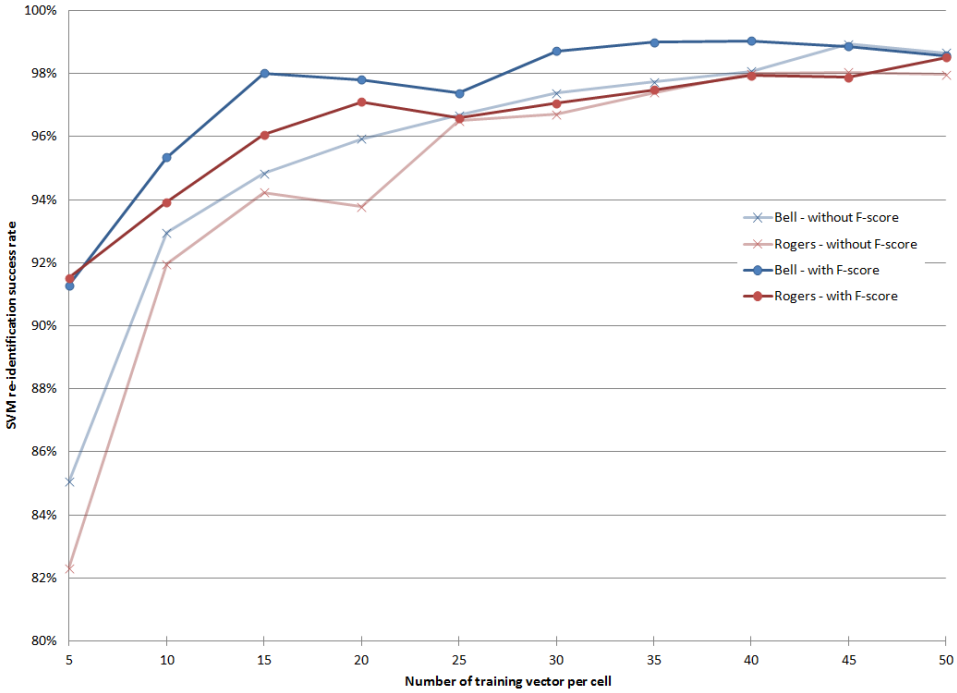



Figure 4.14: Comparing prediction accuracy with and without F-score feature scoring and weighting.

training phase.

Figure 4.14 compares the best combinations of optimization, signal quality filtering and randomness with F-score scoring and weighting and without, and shows that re-identification accuracy converges faster towards high values when using the F-score algorithm, particularly for a small number of feature vectors $n$. Re-identification accuracy is consistently slightly better than in the original case, always in excess of 90% and peaking at 98.5% and 99.0% for Rogers and Bell, respectively. Error bars are not shown in this figure, for clarity.

### 4.2.7 mRMR Feature Scoring and Weighting

The minimum-redundancy-maximum-relevance (mRMR) program from [44] was used against the dataset to determine which feature were the most relevant and the least redundant to use for re-identification. In order to use the mRMR program, however, the trace results had to be significantly altered: the fields need to be comma-separated, contrary to *libsvm* which requires spaces. More importantly was the consideration

that mRMR does not handle large integer as class labels. Since class labels were otherwise constructed to be six-digit numerical values including the Group Cell ID and the Cell ID, class names first had to be normalized from 1 to $l$ by *analyze.sh*.

Feature scoring using the mRMR algorithm resulted in the ranking found in Table 4.4. In this section, feature weighting was accomplished by multiplying each vector element by the corresponding feature score (weighting algorithm #1), although *weight.awk* supported other weighting algorithms. The investigation of re-identification accuracy as a function of the $n$ best features, varying $n$ from 1 to 13, and excluding other features, has not been conducted and is recommended for future work.

Table 4.4: Sample feature scoring results using the mRMR algorithm. See Section 3.1 for details.

| Feature number | Score | Feature description |
|---|---|---|
| 4 | 1.244 | RSTP |
| 5 | 0.594 | OSTP |
| 13 | 0.571 | I & Q Offset |
| 9 | 0.529 | SYNC Correlation |
| 3 | 0.448 | RSEVM |
| 6 | 0.435 | RSRP |
| 1 | 0.392 | EVM |
| 8 | 0.241 | Frequency Error |
| 2 | 0.183 | EVM Peak |
| 7 | 0.089 | RSRQ |
| 12 | 0.074 | Time Offset |
| 11 | 0.026 | Symbol Clock Error |
| 10 | 0.014 | CTE |

Figure 4.15 shows prediction accuracy when the mRMR feature selection algorithm is used, and vector parameters are weighted linearly with the feature score output from the mRMR algorithm, in line with expectations. In Figure 4.15, SVM parameters $C$ and $\gamma$ were optimized by grid search. The anomaly can perhaps be

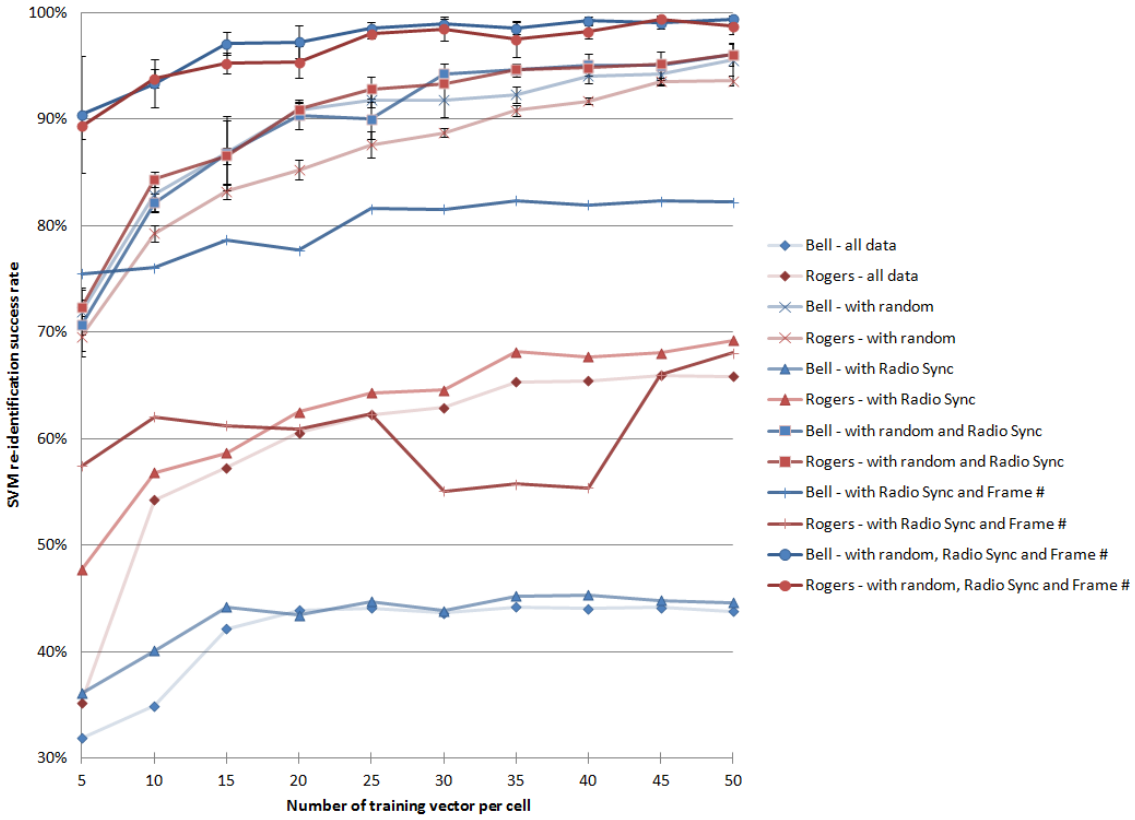



Figure 4.15: Prediction accuracy as a function of the number of training vectors, randomness and signal filtering, with mRMR feature scoring and weighting.

explained by the lack of randomness, suspected to cause the SVM model to optimize for a specific set of feature vectors which ill-represent the radiometric identity of cells when all trace extracts are considered at random. Again, it is shown that prediction accuracy improves under the most stringent signal quality filtering, using radio synchronization and decoded frame number, and using a large number of feature vectors during the training phase.

Figure 4.16 compares the best combinations of optimization, signal quality filtering and randomness with mRMR scoring and weighting and without, and shows that re-identification accuracy converges faster towards high values when using the mRMR algorithm, particularly for a small number of feature vectors $n$. Re-identification accuracy is consistently slightly better than in the original case, always in excess of 89% and peaking at 99.4% for both service providers, respectively. Error bars are not shown in this figure, for clarity. Section 4.2.9 compares the accuracy of F-score and

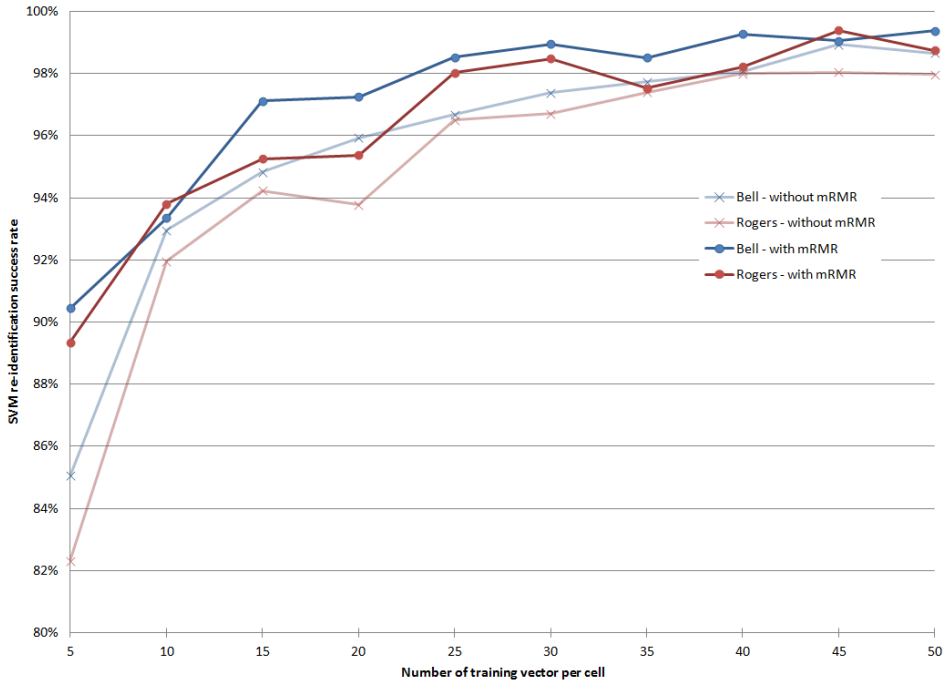



Figure 4.16: Comparing prediction accuracy with and without mRMR feature scoring and weighting.

mRMR using multiple weighting algorithms.

### 4.2.8 Feature Weighting Algorithms

A total of five weighting algorithms were trialed and compared. Section 3.5 discusses the weighting algorithms in detail. The *weight.awk* script was responsible to correctly apply feature weight, based on the command-line selection of weighting algorithm (between 1 and 5), and a standardized feature rank and score file produced by the scoring algorithm (either F-score or mRMR).

Figure 4.17 shows the re-identification accuracy using SVM parameter optimization by grid search, the strictest signal quality filtering and randomness, for five weighting algorithms, using the F-score algorithm for feature scoring. As a reference, the non-scored and non-weighted curves are also shown, in black. Note that the fifth weighting algorithm is not using any feature scoring and is a special case closely resembling the features chosen by Brik et al. in [2]. It can be seen that all weighting algorithms result in some accuracy improvement, with the exception of $w(a_i) = 2^{F(a_i)}$

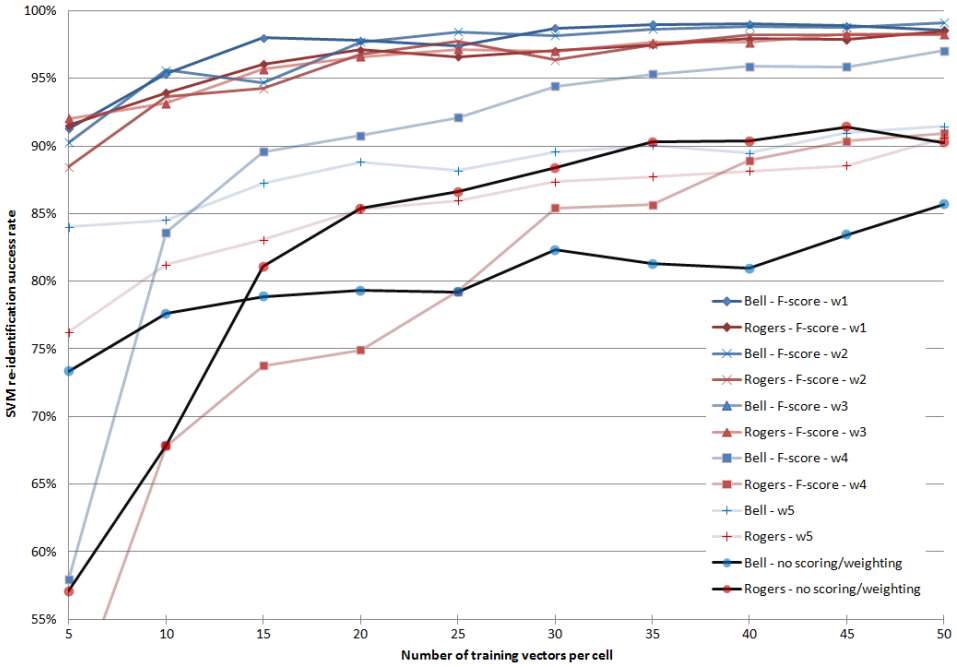



Figure 4.17: Prediction accuracy as a function of the number of training vectors and weighting algorithm, with randomness, SVM parameter optimization and signal filtering, using the F-score scoring algorithm.

with less than 25 feature vectors used during SVM training, for one of the service provider. Error bars are not shown in this figure, for clarity.

Figure 4.18 shows the re-identification accuracy using SVM parameter optimization by grid search, the strictest signal quality filtering and randomness, for five weighting algorithms, using the mRMR algorithm for feature scoring. As a reference, the non-scored and non-weighted curves are also shown, in black. Note that the fifth weighting algorithm is not using any feature scoring and is a special case closely resembling the features chosen by Brik et al. in [2], and is as such the same as in the previous section. It can be seen that all weighting algorithms result in some accuracy improvement. Error bars are not shown in this figure, for clarity.

### 4.2.9 Comparing F-score against mRMR

Both feature scoring algorithms provided marginal improvements in re-identification success rates, but provided faster convergence to high accuracy under a small number of training vectors. Figure 4.19 compares both feature scoring algorithms, whilst using

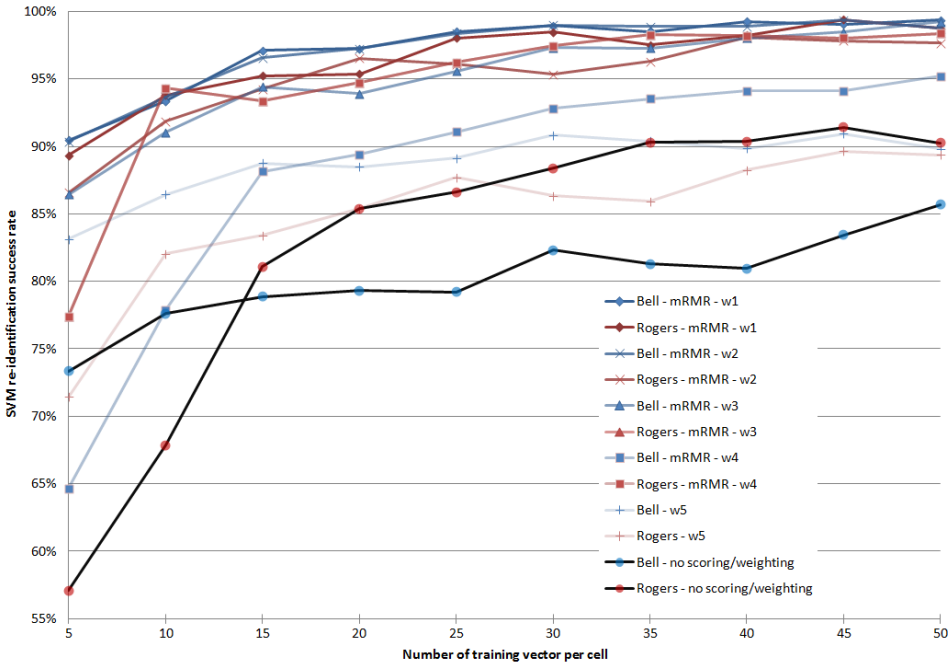



Figure 4.18: Prediction accuracy as a function of the number of training vectors and weighting algorithm, with randomness, SVM parameter optimization and signal filtering, using the mRMR scoring algorithm.

SVM parameter optimization by grid search, the strictest signal quality filtering and randomness. Only two of five weighting algorithms are represented in this figure (linear weight by score or by rank). As a reference, the non-scored and non-weighted curves are also shown, in black, as well as the Brik's special case with only four features considered ($w5$). It can be seen that both feature scoring algorithm results in marginal improvements in re-identification accuracy, particularly when a small number of feature vectors are considered for the SVM model during the training phase, and that the differences between scoring algorithms and weighting algorithms #1 and #2 are not significant. Interestingly, results obtain under feature selection akin Brik et al. were notably below scored and weighted combinations, but still above the original case that considers all feature parameters. This highlights an important conclusion which indicates it is best not to use all 13 features when not using feature scoring and weighting, as it appears some features negatively impact re-identification. Error bars are not shown in this figure, for clarity.

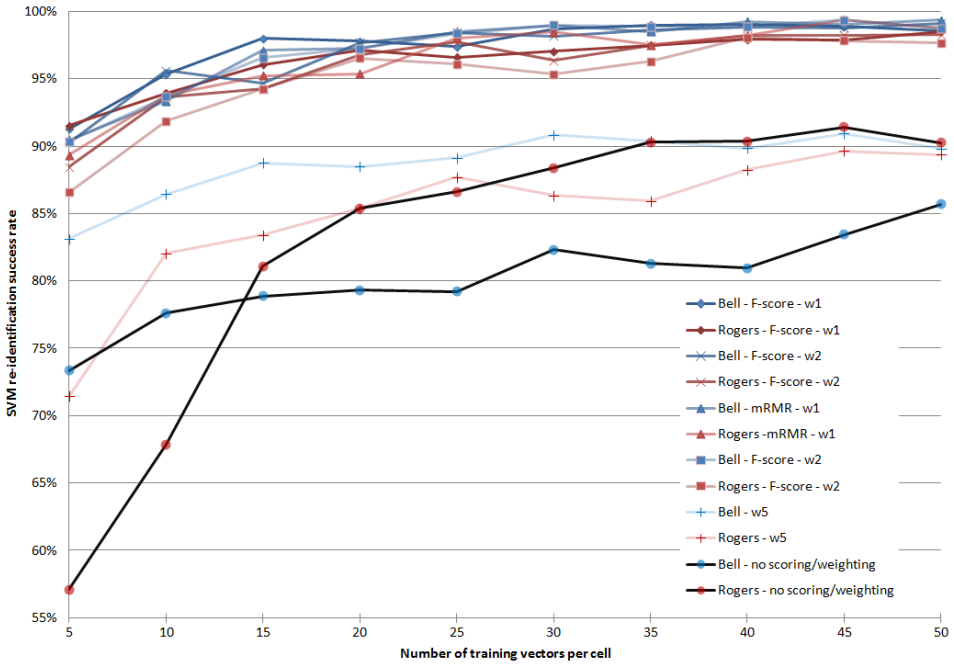



Figure 4.19: Prediction accuracy, F-score vs mRMR, as a function of the number of training vectors, scoring and weighting algorithms, with randomness, SVM parameter optimization, signal filtering.

### 4.2.10 Analysis of the Impact of Adverse SNR on Re-identification

Although Section 4.2.1 discusses how low signal reception quality negatively impacts re-identification, comparing the results when filtering for synchronization status and decoded frame number is enabled or disabled, it is also interesting to investigate if re-identification is possible for the case where the SVM training model is conducted with feature vectors extracted under high SNR, whereas prediction is attempted using feature vectors extracted under low SNR. For this experiment, only feature vectors extracted under the best signal conditions were used to build a set of $n$ training vectors. Subsequently, all feature vectors matching one of the possible cell, but having a poor signal quality (no frame number decoded) were subjected to SVM re-identification.

Figure 4.20 shows the re-identification success rates for feature vectors extracted without LTE frame number decoded nor synchronization. As a reference, the non-scored and non-weighted curves are also shown, in black. It shows that results are generally significantly lower than the general case, and hover between 50% and 60%. This is an indication that RF channel conditions adversely impact re-identification,

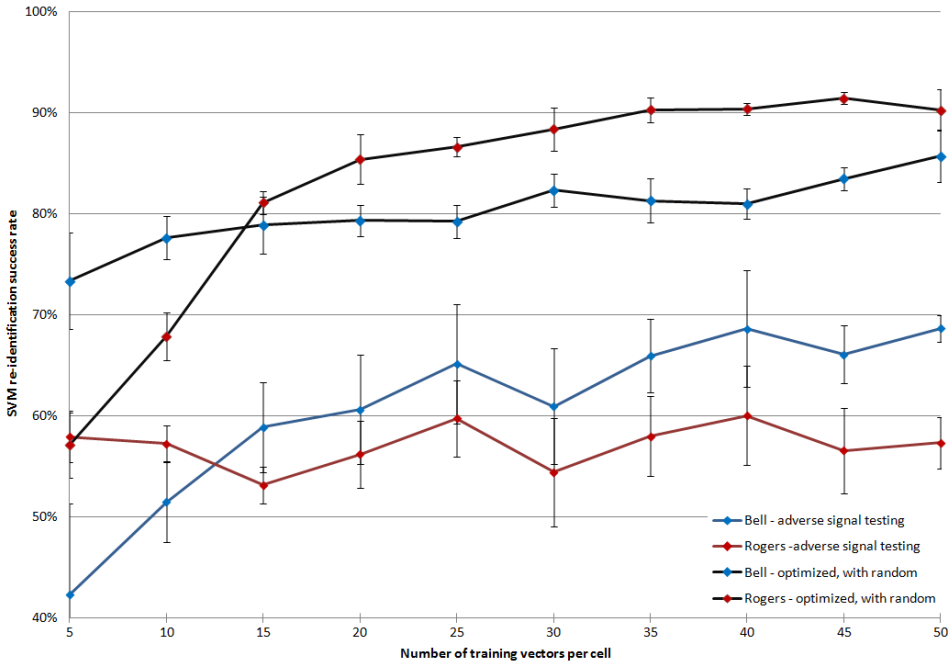



Figure 4.20: Prediction accuracy in low SNR conditions, as a function of the number of training vectors, scoring and weighting algorithms, with randomness, SVM parameter optimization, signal filtering.

in line with expectations. However, the Agilent VSA documentation warned against considering any measurements captured without synchronization. As such, any practical application should at the very least attempt to reject feature vectors extracted without synchronization. A more thorough study of the effects of the RF channel was not possible with the dataset collected, because there was not a sufficient number of cells that were observed with high SNR from multiple recording positions to create an interesting re-identification problem (7 and 8 cells were observed from multiple recording positions, with the strictest signal quality filtering, however only 3 and 4 cells had a sufficient number of feature vectors for analysis).

### 4.2.11 Combining Cellular Providers

Combining trace data from two cellular providers is interesting in that it increases the cell population and should render the re-identification more difficult. However, in the current implementation, cell IDs used as class labels are computed by combining the Group Cell ID and the Cell ID from the VSA software into a 6 digit integer.

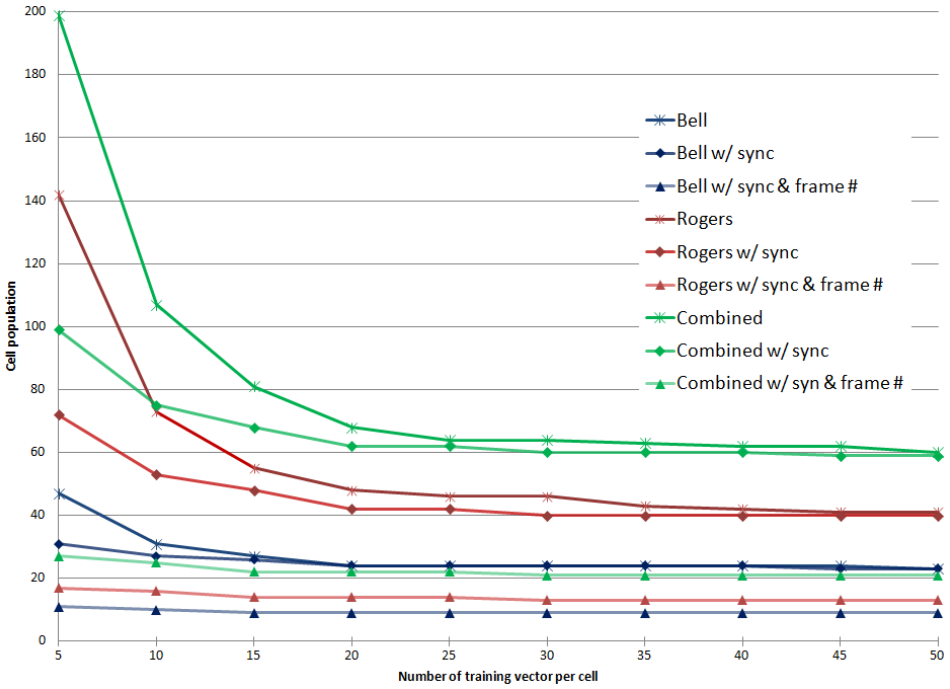



Figure 4.21: Cell population as a function of the training bin size, when combining network operators.

It was noted that a number of cell IDs are used by each provided to identify cells on their respective networks. As such, collapsing the data from both providers has the effect of confusing the model for a number of cells: vectors from two different radio entities are labelled with the same class ID. The phenomenon is quantified in Table 4.1, where at most 123 cell IDs (no filtering) or at least 2 cell IDs (radio synchronization and decoded frame number) use conflicting cell IDs. Figure 4.21 shows the cell population of combined providers as a function of the minimum number of feature vectors retained for the SVM training model, as well as the cell populations of either provider when considered independently.

This experiment also has little practical application, since the two providers are using different frequency bands, it will always be possible to discriminate between them based on the allocated frequencies. It would be possible to eliminate the conflicts by computing cell IDs differently e.g. inserting a $7^{\text{th}}$ digit and use this $7^{\text{th}}$ digit to identify the provider. However this approach would have little benefit since it effectively would not collapse the traces of both providers.

Nonetheless, the experiment showed that re-identification peaked at 98.4% and

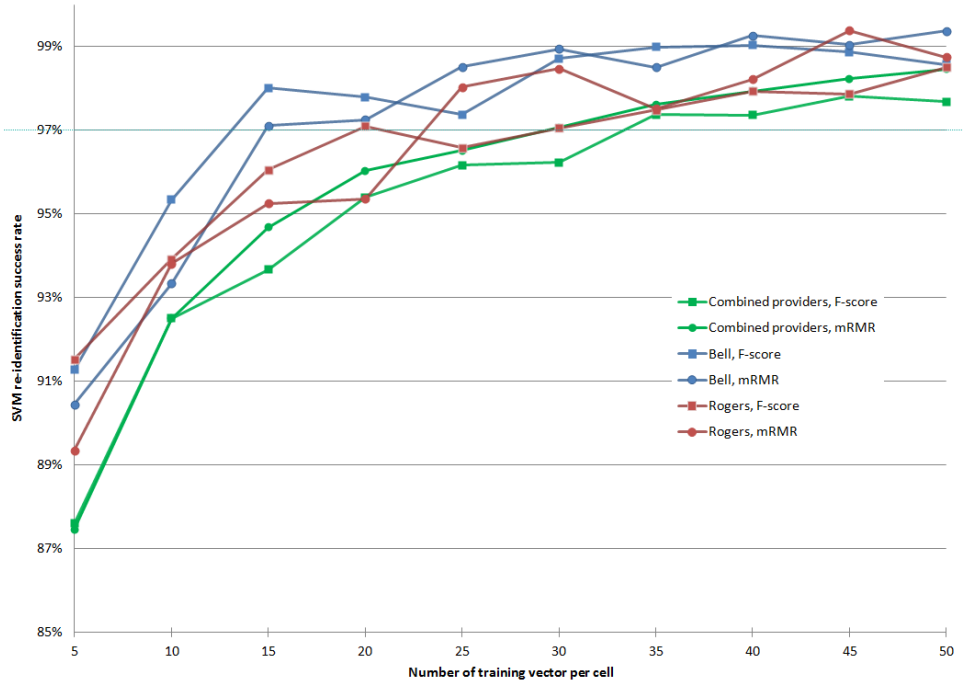



Figure 4.22: Prediction accuracy as a function of the number of training vectors, scoring and weighting algorithms, with randomness, SVM parameter optimization, signal filtering, with combined network operators.

97.8% using the best signal quality filtering and kernel parameter optimization, for mRMR and F-score feature scoring algorithms, respectively. Figure 4.22 shows that combining the traces for both providers has a negative impact on re-identification accuracy, which can likely be explained that some radiometric properties assigned to a single class actually originate from two distinct transmitters sharing the same cell ID, or class label. Error bars are not shown in this figure, for clarity.

## 4.3 Analysis and Recommendations

The experimental results show that with some conditioning, near perfect re-identification accuracy can be achieved for LTE transmitters belonging to either of the two cellular providers. In order to maximize SVM prediction accuracy, it is recommended that only vectors corresponding to high signal quality be considered during both the training and testing phases. It is further recommended that the dataset be first randomized and the optimal SVM parameters $C$ and $\gamma$ be discovered using grid

search, both recommended by Hsu in [35]. The use of feature scoring and weighting always resulted in faster convergence with a small number of training vectors and results in non-negligible improvements in re-identification accuracy. The choice of the weighting algorithm or the feature scoring algorithm did not however result in significant variations, indicating that as long as key features are given more importance, improvements are obtained, regardless of the relative importance between features. Both feature scoring algorithms were found to score features in a comparable order, even though they operated fundamentally differently, in that mRMR also attempted to reduce redundancy between features; three out of the top five features were common between the two scoring algorithms. Interestingly, a feature selection of the four features closest to the Brik et al. experiments in [2] also resulted in improvements over the general non-optimized case, indicating that some of the features are detrimental to re-identification when considered in the SVM training model, particularly when the SVM parameters are not optimized by grid search. Lastly, the prediction accuracy was shown to improve when a larger set of feature vectors are supplied to build the SVM training model (best results consistently obtained with 45 to 50 training vectors); however, when feature scoring and weighting is applied, and optimal $C$ and $\gamma$ are discovered, results well above 90% are obtained with as few as 10 training vectors. Larger sets of training vector in those experiments only result in marginal accuracy gains. It should be noted that in this thesis, the re-identification of each unknown emitters was conducted using a single feature vector (i.e. a testing bin size of 1), leading to great results. In practical applications, it may be possible to gather several feature vectors from an unknown transmitter in order to determine its identity, which is expected to result in even greater accuracy.

# Chapter 5

# Conclusions and Future Work

LTE is an emerging communication standard already available in many urban centres worldwide. Competing 4G cellular technologies have been abandoned by standard bodies (e.g. UMB) or by the majority of network operators (e.g. WiMAX), opening the door for a first truly global cellular standard. This research demonstrated that sufficient differences exist between LTE transmitters to enable near-perfect re-identification, for both service providers, based on radiometric properties of the downlink signal.

This chapter first reviews the results achieved, in line with the contributions listed in Chapter 1. Potential applications of these excellent results in solving practical problems are also proposed. Next, limitations observed during this research are highlighted, followed by several recommendations for further research, building upon the results shown herein.

## 5.1 Contributions, Results and Applications

This section first reviews the research objectives considered at the onset of this study, and presents a brief overview of the results achieved. Practical applications that could use radiometric identifications are then presented.

### 5.1.1 Main Contributions and Results Overview

This research activity attempted to and succeeded in presenting novel contributions in the following areas:

- First successful radiometric identification of LTE transmitters. Re-identification

of LTE downlink transmitters was shown to be possible with an accuracy in excess of 99%, using commercially-available hardware and software, combined with an established classifier algorithm (SVM).

- First successful use of a weighted classifier algorithm (weighted-SVM) to conduct radiometric identification. This thesis showed that feature scoring and weighting had a beneficial impact on re-identification due to faster convergence towards accurate prediction with fewer training vectors. The F-score and mRMR scoring algorithms scored features in a comparable fashion, as expected. Four weighting algorithms (linear by score or rank, exponential by score or rank) were trialed for both F-score and mRMR scoring, as well as a special case of feature selection using only four features. All weighting algorithms resulted in similar improvements, whereas the feature selection trial exhibited lower improvement but still outperformed the general non-optimized case.

- First examination of the impact of variable modulation schemes on radiometric identification. This thesis demonstrated that radiometric identification was successful in spite of the LTE downlink supporting multiple modulation schemes. It is not certain, however, that all modulation schemes were present in the recordings analyzed by the VSA software, because the *Error Summary* table did not specifically provide error summary values for each of the modulation observed in the downlink, against expectations (see Figure 3.4). It is believed that the various modulation schemes observed on the different downlink channels were aggregated in a single set of error values and this was shown not impede accurate re-identification.

### 5.1.2 Applications

#### Enhancement to Identity Verification and Authentication

The LTE standard has been designed from the ground up with security and privacy in mind. As such, there are a number of measures and techniques already in place to protect the identity of users, and ensure communication and location privacy.

Radiometric identification can be considered as a complement to existing cryptographic techniques and should be fairly easy to implement. Radio receiving equipment is already capable of performing some analysis of the modulation characteristics of

nearby emitters, providing network operators the ability to black-list handsets which fail to comply to communication standards. Based on the results of this research, it should be possible to validate the identity of a base station using its radiometric properties, in order to detect man-in-the-middle attacks or spoofing that might have gone undetected otherwise. Expensive and highly sensitive spectrum analysis equipment such as the one used for this thesis may not be required, as demonstrated by work by Edman and Yener in [33] showing re-identification success rates of 87% using inexpensive equipment such as the USRP. It is believe that radiometric identification opportunities using band-optimized radio electronics present in modern phones is conceivable in the near future.

In a conceptual system, network operators could be responsible to maintain a valid classifier model which describes the equipment they legitimately field. This could be done by the network operator or crowd-sourced by users running a trusted application which would submit feature vectors over time. During testing, the UE could query the provider with a test feature vector, or could run the classifier locally if a valid model file had been downloaded ahead of time.

#### User Tracking

The application of this technique could be applied to UE, such that it may be possible to track LTE users using radiometric properties of their device without having to decipher private user identification strings. The recognition of their radiometric identity alone could allow an attacker to track a mobile user in spite of changing user identifiers provided by the network for privacy protection.

## 5.2 Limitations and Recommendations

A number of recommendations flow from the observations of the results discussed in Chapter 4.

- It is recommended that in future research or practical applications, only vectors collected in the presence of radio synchronization and a decoded LTE frame number be considered for the generation of the training model or during the testing phase. It is further recommended that feature scoring and weighting be consistently applied as it results in significant accuracy improvements, particularly with fewer training vectors.

- In this research, the population of candidate cells eligible for re-identification shrunk with larger minimal numbers of training vectors. This effect can be overcome by extracting significantly more feature vectors from the RF recordings. Although the RF recordings captured during the experiment were sufficiently long, the time taken to step through and extract feature vectors did not permit the extraction of additional feature vectors. Experimental results have shown that training bins in excess of 35 vectors result in marginal improvements.

- All results obtained without first randomizing the set of feature vectors resulted in important anomalies and significantly lower re-identification accuracy. As such, it is recommended that all experiments and practical applications apply randomness to the set of feature vectors to maximize the accuracy of the SVM training model. This may be difficult to achieve in a practical applications.

- All experiments conducted in this thesis consisted in RF recordings of static emitters from fixed positions. Practical applications will most certainly involve a mobile emitter or radiometric identification from a mobile platform, or both. The effect of emitter or identifier motion should be further studied.

- All results were obtained using recordings made on a single day. It is not clear that training models created from RF recordings on a given day would result in effective radiometric identification on a later day, due to component aging and environmental conditions affecting the elements of the transmitter RF path. Furthermore, it is possible that some aspects of the environment and radio channel be mitigated by building a training model which considers vectors gathered during multiple recordings, separated by a few hours or a few days.

- The overhead of maintaining UE radiometric identities should be considered, in particular when the study of feature aging has answered the question concerning the longevity of feature vectors, and the performance of classifiers faced with a very large population of emitters.

## 5.3 Future Work

As LTE is nearing global availability with unprecedented fielding rates, it is anticipated that research opportunities will continue to be relevant for years to come.

Promising subsequent research topics are presented, which build upon the results achieved and limitations identified by this study.

### 5.3.1 Effects of the RF Channel on Re-identification

Radiometric identification in the modulation domain is said to be more resilient to changing RF channel conditions. However, many of the features extracted by the VSA reflect properties that can be affected by the received signal strength. This correlation needs to be further studied to determine if re-identification can remain effective in changing RF channels, particularly in light of the results obtained using a subset of the 13 features present in the feature vectors, which outperformed the general non-optimized case.

### 5.3.2 Radiometric Identification of User Equipment

This research activity focused on the radiometric identification of LTE base stations. Successful radiometric identification of UEs could broaden the practical applications of this research, as discussed above. In LTE, UE transmissions are multiplexed using SC-FDMA, which has not yet been studied for the purpose of radiometric identification. The accuracy of an SVM classifier against a very large set of emitters is also an important consideration, as is the suitability of SVM classifiers against an ever-changing population of candidates (e.g. as new UEs are sold to consumers, a new classifier model would have to be generated and distributed).

### 5.3.3 Other Transmitter Error Information

The VSA software provides a number of traces which could lead to features enhancing radiometric identification of emitters. The *Error Summary* table was exclusively used in this research activity. Other traces may be worth investigating and may lead to equal or better performance. Notably, the *Error Vector Spectrum* trace, which provides error information per subcarrier, the *Error Vector Time* trace, which provides error per symbol time, the *CTE* and *MIMO CTE* traces, and the *Frequency Error Per Slot* trace are all likely capable of providing enriched feature vectors useful in improving emitter identification.

### 5.3.4 Emitters in Motion

Particularly useful for the study of UE radiometric identification is the notion of signature quality for emitters in motion. The research activity proposed in this document focused on static LTE eNB and all recordings were conducted from a fixed position. Similar experiments could be conducted against LTE eNBs from a mobile platform, or conversely, RF recordings of mobile emitters such as LTE UEs could be captured. In either case, the impact on re-identification accuracy should be considered when the training set, the testing set, or both sets, are computed from RF recordings with mobility.

### 5.3.5 Component Aging and Temperature

Due to the logistics involved in recordings live signals from the City of Ottawa, using borrowed DRDC equipment, it was not possible to return and gather additional live recordings to study the stability of emitter parameters over time. Of importance is the determination whether electronic components in the RF path of the emitters are subject to wear and tear due to usage and environmental factors. Once the feature aging is quantified, studying its effect on re-identification accuracy is important for most of the practical applications suggested. The recommended approach consists in training the classifier using feature vectors extracted from RF recordings gathered at time $t_1$ and query the classifier model with unknown emitter feature vectors extracted from RF recordings conducted at time $t_2$, where $t_2 >> t_1$. Similarly, Polak et al. discuss in [1] that power amplifiers often do not respond linearly with temperature. This may cause radiometric identification to under-perform if the feature vectors used in the training phase were extracted at a temperature significantly different from the temperature at which test feature vectors are extracted.

### 5.3.6 Alternate Classifier Algorithms

SVMs have been shown by Brik et al. to be particularly effective at classifying emitters based on parameters obtained from the analysis in the modulation domain [2]. Other classifiers have been examined, namely $k$-nearest neighbours (kNN) in [2] and probabilistic neural network (PNN) [27, 50], however the issue of scalability (memory requirement per profile) prohibited its use in real time systems [4].

Of particular interest is the random forests algorithm, presented by Breiman in

[51]. Random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forests, each node is split using the best among a sub-set of predictors randomly chosen at that node. This somewhat counter-intuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, SVMs and neural networks, and is robust against overfitting [52].

The WEKA machine learning environment proposed by Hall et al. in [49] seems well suited for this study as it provides a single environment in which several machine learning algorithms can be applied against the same dataset, including SVM and random forests.

### 5.3.7 Principal Component Analysis (PCA)

As a substitute for feature scoring, it may be interesting to investigate the performance of PCA coefficients through the classifier algorithm. PCA, also known as Karhunen-Loève transformation (KLT), is a mathematical technique which converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. One distinction between a scoring algorithm such as mRMR and PCA is that multiple features can be combined into a single principal component. Both techniques should return comparable results as they both attempt to identify highly-correlated features. It is conceivable that principal components may not be equally relevant to accurate re-identification, and a second stage of component scoring and weighting, instead of feature scoring and weighting may yield the best results.

### 5.3.8 Feature Exclusion

The feature scoring algorithms presented herein ranked the 13 features in a comparable fashion. It may be possible to exclude a subset of the low-ranking features and still achieve acceptable re-identification accuracy whilst reducing the classification problem's complexity. Efforts in this direction may help develop practical and cost-effective applications.

# List of References


[1] A. Polak, S. Dolatshahi, and D. Goeckel, "Identifying wireless users via transmitter imperfections," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 7, pp. 1469–1479, 2011.

[2] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless device identification with radiometric signatures," in *Proc. 14th ACM Int. Conf. Mobile Computing and Networking*. ACM, 2008, pp. 116–127.

[3] J. Hall, M. Barbeau, and E. Kranakis, "Detection of transient in radio frequency fingerprinting using signal phase," *Wireless and Optical Communications*, pp. 13–18, 2003. [Online]. Available: http://pdf.aminer.org/000/274/667/robust_detection_of_signals_with_unknown_frequency_and_phase.pdf

[4] ——, "Enhancing intrusion detection in wireless networks using radio frequency fingerprinting," in *Proc. 3rd Int. Conf. on Communications, Internet and Information Technology (IASTED)*, 2004, pp. 201–206.

[5] M. Barbeau, J. Hall, and E. Kranakis, "Detecting impersonation attacks in future wireless and mobile networks," *Secure Mobile Ad-hoc Networks and Sensors*, pp. 80–95, 2006. [Online]. Available: http://people.scs.carleton.ca/~canccom/Publications/at_2006.pdf

[6] J. Hall, M. Barbeau, and E. Kranakis, "Radio frequency fingerprinting for intrusion detection in wireless networks," July 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.6584&rep=rep1&type=pdf

[7] M. Riezenman, "Cellular security: better, but foes still lurk," *Spectrum, IEEE*, vol. 37, no. 6, pp. 39 –42, jun 2000.

[8] M. Shin, J. Ma, A. Mishra, and W. Arbaugh, "Wireless network security and interworking," *Proc. IEEE*, vol. 94, no. 2, pp. 455–466, 2006.

[9] U. Meyer and S. Wetzel, "A man-in-the-middle attack on UMTS," in *Proceedings of the 3rd ACM workshop on Wireless security*. ACM, 2004, pp. 90–97.

[10] H. Rohling, Ed., *OFDM Concepts for Future Communication Systems*. Springer-Verlag, 2011.

[11] M. Sauter, *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. Wiley, 2011.

[12] F. Khan, *LTE for 4G mobile broadband: air interface technologies and performance*. Cambridge University Press, 2009.

[13] M. Rumney, Ed., *LTE and the Evolution of 4G Wireless Design and Measurement Challenges*. Agilent Technologies, 2009.

[14] E. Dahlman and J. Parkvall, S.and Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. Elsevier, 2011.

[15] *IEEE Standard for Information Technology-Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std. IEEE Std 802.11-2007, 2007, revision of IEEE Std 802.11-1999.

[16] J. Liu, J. Lee, L. Li, Z.-Q. Luo, and K. Wong, "Online clustering algorithms for radar emitter classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1185–1196, 2005.

[17] J. Matuszewski, "Specific emitter identification," in *International Radar Symp.*, 2008, pp. 1–4.

[18] S. Dolatshahi, A. Polak, and D. L. Goeckel, "Identification of wireless users via power amplifier imperfections," in *Rec. 44th ASILOMAR Conf. Signals, Systems and Computers*, 2010, pp. 1553–1557.

[19] A. Polak and D. Goeckel, "RF fingerprinting of users who actively mask their identities with artificial distortion," in *Rec. 45th ASILOMAR Conf. Signals, Systems and Computers*, 2011, pp. 270–274.

[20] Z. Li, W. Xu, R. Miller, and W. Trappe, "Securing wireless systems via lower layer enforcements," in *ACM Workshop Wireless security*. ACM, 2006, pp. 33–42.

[21] G. Chandrasekaran, J.-A. Francisco, V. Ganapathy, M. Gruteser, and W. Trappe, "Detecting identity spoofs in ieee 802.11e wireless networks," in *Proc. IEEE Global Telecommunications Conf. (GLOBECOM)*, 2009, pp. 1–6.

[22] T. Chen, W. Jin, and J. Li, "Feature extraction using surrounding-line integral bispectrum for radar emitter signal," in *IEEE International Joint Conf. Neural Networks*. IEEE, 2008, pp. 294–298.

[23] I. Kennedy, P. Scanlon, F. Mullany, M. Buddhikot, K. Nolan, and T. Rondeau, "Radio transmitter fingerprinting: A steady state frequency domain approach," in *IEEE 68th Vehicular Technology Conf.*, 2008, pp. 1–5.

[24] G. O. Zamora, S. Bergin, and I. Kennedy, *Novel Algorithms and Techniques in Telecommunications and Networking*. Springer Netherlands, 2010, ch. Using Support Vector Machines for Passive Steady State RF Fingerprinting, pp. 183–188. [Online]. Available: http://dx.doi.org/10.1007/978-90-481-3662-9_31

[25] B. Danev, A. de Splinder, H. Luecken, and S. Capkun, "Physical-layer identification: Secure or not?" ETH, Department of Computer Science, http://dx.doi.org/10.3929/ethz-a-006836297, Tech. Rep. 637, 2009. [Online]. Available: ftp://ftp.inf.ethz.ch/pub/publications/tech-reports/6xx/637.pdf

[26] B. Danev and S. Capkun, "Transient-based identification of wireless sensor nodes," in *Int. Conf. Information Processing in Sensor Networks*, 2009, pp. 25–36.

[27] D. Shaw and W. Kinsner, "Multifractal modelling of radio transmitter transients for classification," in *Conf. Proc. WESCANEX '97: Communications, Power and Computing, IEEE*, 1997, pp. 306–312.

[28] O. Ureten and N. Serinken, "Detection of radio transmitter turn-on transients," *Electronics Letters*, vol. 35, no. 23, pp. 1996–1997, 1999.

[29] ——, "Bayesian detection of Wi-Fi transmitter RF fingerprints," *Electronics Letters*, vol. 41, no. 6, pp. 373–374, 2005.

[30] A. Kawalec and R. Owczarek, "Radar emitter recognition using intrapulse data," in *Int. Conf. Microwaves, Radar and Wireless Communications (MIKON)*, vol. 2, 2004, pp. 435–438.

[31] I. Kennedy and A. Kuzminskiy, "RF fingerprint detection in a wireless multipath channel," in *Int. Symposium on Wireless Communication Systems (ISWCS)*, 2010, pp. 820–823.

[32] Y. Shi and M. Jensen, "Improved radiometric identification of wireless devices using MIMO transmission," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 4, pp. 1346 –1354, 2011.

[33] M. Edman and B. Yener, "Active attacks against modulation-based radiometric identification," *Rensselaer Institute of Technology, Technical report*, pp. 09–02, 2009.

[34] C. Campbell and Y. Ying, *Learning with Support Vector Machines*, R. Brachman and T. Dietterich, Eds. Morgan & Claypool Publishers, 2011.

[35] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," April 2010. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[36] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. [Online]. Available: http://dx.doi.org/10.1007/BF00994018

[37] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[38] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[39] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction*. Springer, 2006, pp. 315–324.

[40] A. Kawalec and R. Owczarek, "Specific emitter identification using intrapulse data," in *1st European Radar Conference (EURAD)*, 2004, pp. 249–252.

[41] Y. Chen, "Feature selection tool for libsvm," September 2012, retreived 16 Sep 2012. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

[42] A. Technologies. (2012, July) Agilent 89600 VSA Software, v.15.0. [Online]. Available: http://www.home.agilent.com/en/pc-1905089/89600-vsa-software?&cc=CA&lc=eng

[43] *Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) conformance testing*, 3GPP Std. TS 36.141, Rev. 8.12.0, 07 2012. [Online]. Available: http://www.etsi.org/deliver/etsi_ts/136100_136199/136141/08.12.00_60/ts_136141v081200p.pdf

[44] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[45] A. Schneider, "GPS visualizer," May 2012. [Online]. Available: http://www.gpsvisualizer.com/

[46] F. Demers, "Data gathering 13 Feb 2012," 2012. [Online]. Available: http://www.fdemers.com/LTE

[47] Y. Chen and C. Lin, *Feature Extraction, Foundations and Applications*. Springer, 2006, ch. Combining SVMs with various feature selection strategies, pp. 315–324.

[48] S. Punter, "Southern ontario cell phone page," 2012. [Online]. Available: http://www.arcx.com/sites/

[49] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[50] A. Hunter, "Feature selection using probabilistic neural networks," *Neural Computing & Applications*, vol. 9, no. 2, pp. 124–132, 2000. [Online]. Available: http://eprints.lincoln.ac.uk/1913/1/FeatureSelectionGAPNN.pdf

[51] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://leg.ufpr.br/lib/exe/fetch.php/wiki:internas:biblioteca:randomforests.pdf

[52] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://www.webchem.science.ru.nl:8080/PRiNS/rF.pdf