



University
of Cyprus

Department of Electrical and Computer Engineering

An Algorithm Agnostic Framework for the Evaluation and
Learning of Robust Classifiers for Data under Uncertainty

Elisavet Charalambous

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the University of Cyprus

December , 2019

© Elisavet Charalambous, 2019

APPROVAL PAGE

Elisavet Charalambous

An Algorithm Agnostic Framework for the Evaluation and Learning of Robust
Classifiers for Data under Uncertainty

The present Doctorate Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering, and was approved on December 16, 2019 by the members of the Examination Committee.

Committee Chair _____
Prof. Theocharis Theocharides

Research Supervisor _____
Prof. Constantinos Pitris

Committee Member _____
Prof. Stelios Timotheou

Committee Member _____
Prof. Vasiliki Kassianidou

Committee Member _____
Prof. Michalis Michaelides

DECLARATION OF DOCTORAL CANDIDATE

The present doctoral dissertation was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

Signature

Elisavet Charalambous

Abstract

Machine Learning (ML) has become in the recent years increasingly ubiquitous for its classification and clustering capabilities, with a wide range of applications in science, engineering, social sciences and humanities, including archaeology and security.

An algorithm's effectiveness in correctly classifying samples to the desired class is influenced by factors such as its intrinsic characteristics and parametrization, training and evaluation methods as well as the appropriateness of the input dataset. As a result, an algorithm's performance may be greatly influenced with variability of any of these factors; when it has not been previously considered. Nowadays, ML techniques find applicability to countless domains towards the resolution of problems that range from very simple to very complex; these usually rely on patterns and inference.

It has become rather important to develop methods and metrics, algorithmic agnostic, that allow estimating a models' ability of consistently producing acceptable results; a practice that is non trivial. In this Thesis, we propose an algorithmic agnostic methodology for learning robust classifiers for data with uncertainties. The proposed methodology is agnostic of the selected classification method and emerges as a result of thorough analysis of factors that influence the classification result and emerge from factors related to the application domain and dataset characteristics. The developed design follows a systematic approach and well-established methods, such as bootstrapping with replacement and the 5x2 cross validation (paired t-test and F-test) tests, to ensure the results are statistically significant.

The produced results indicate that the evaluation of robustness in classification is possible, while investigation of inter-class relationships on classification results may provide expert researchers with additional information for data samples with low classification confidence.

The suggested methodology has been validated against two case studies: (a) classification of scarce chemical compositional archaeological data from ceramics and, (b)

classification of audio samples for acoustic event detection in the field of intelligent surveillance for security purposes. Finally, an open source web-based tool realising the proposed framework is presented for use by other scientists and application domain experts.

Acknowledgments

Firstly I would like to thank my current advisor Prof. Konstantinos Pitris and Prof. Theocharis Theocharides for their guidance and encouragement towards completing this work. It has been a long trip with many ups and downs and without their last mile involvement this Thesis would not have been realised.

Then, I would like to express my sincere gratitude to Dr. Demetrios G. Eliades and Dr. George M. Milis for the continuous support of my Ph.D study and related research. Even though, they did not have the role of my formal advisor, they supported this effort – since day one – with indescribable patience and motivation.

I would also like to thank Prof. Georgios Mitsis who has served as my advisor for an important part of my PhD journey. His immense knowledge and guidance incited me to widen my research from various perspectives.

Besides members of the Electrical and Computer Engineering Department and KIOS Center of Excellence, I would like also express my gratitude to Dr. Maria Dikomitou and Prof. Vasiliki Kassianidou for their support and knowledge with respect to archaeological data analysis. Their endless sources and immense patience assisted greatly in traversing my path towards the implementation of methods for the analysis of archaeological data.

My sincere thanks also goes to Mr. Nikolaos Koutras, Managing Director of ADITESS ltd. for his understanding and support towards the completion of my PhD. My placement as Software Engineer and Researcher in ADITESS for the past five years allowed me to research further and deploy my solutions in practical problems in the area of security and cybercrime.

A very special thanks goes to my dear friend and excellent scientist Dr. Maria Terzi who has been part of my life for the past twelve years. Maria and I started together our undergraduate studies in 2007 in Lancaster University, England, and have been friends ever since. Fate has made it that Maria is now member of the KIOS Research

Center of Excellence and I could have never hoped for a better colleague.

Last but not the least, I would like to thank my family for supporting me spiritually throughout writing this Thesis and my my life in general.

Research in Case Study I was supported by the European Union under the 7th Framework Programme “FP7-PEOPLE-2010-ITN”. Grant agreement number 265010 – “New Archaeological Research Network for. Integrating Approaches to Ancient Material” (NARNIA-ITN).

Publications

Published journal publications

1. E. Charalambous, M. Dikomitou-Eliadou, G. M. Milis, G. Mitsis, and D. G. Eliades, “An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relationships”, *J. Archaeol. Sci. Reports*, vol. 7, pp. 465–471, 2016.
2. E. Tsakalos, J. Christodoulakis, and L. Charalambous, “The Dose Rate Calculator (DRc) for Luminescence and ESR Dating-a Java Application for Dose Rate and Age Determination”, *Archaeometry*, vol. 58, no. 2, pp. 347-352, 2015.
3. J. Arraiza et al., “Fighting volume crime: an intelligent, scalable, and low cost approach”, *J. Polish Saf. Reliab. Assoc.*, vol. 6, no. 3, pp. 1-8, 2015.
4. M. Skitsas, N. Efstathiou, E. Charalambous, N. Koutras, and C. Efthymiou, “Towards the Protection of Critical Information Infrastructures using a Lightweight, Non-intrusive Embedded System”, *J. Polish Saf. Reliab. Assoc.*, vol. 7, no. 1, pp. 187-192, 2016.
5. E. Charalambous, R. Bratskas, N. Koutras, G. Karkas, and A. Anastasiades, “Email forensic tools: A roadmap to email header analysis through a cybercrime use case”, *J. Polish Saf. Reliab. Assoc.*, vol. 7, no. 1, pp. 21–28, 2016.

Published book chapters

1. E. Charalambous, M. Skitsas, N. Efstathiou, and N. Koutras, “A Digital Decision Support System for Efficient Policing in Urban Security in a Community Policing Context,” in *Synergy of Community Policing and Technology*, Springer, 2019, pp. 1–14.

2. N. Kolokotronis, S. Shiaeles, E. Bellini, E. Charalambous, D. Kavallieros, O. Gkotsopoulou, P. Clement, A. Bellini and G. Sargsyan, "Cyber-Trust: The Shield for IoT Cyber-Attacks." Resilience and Hybrid Threats: Security and Integrity for the Digital World, vol.55, pp 76-93, 2019.
3. S. Tsekeridou, G. Leventakis, G. Kokkinis, E. Charalambous, D. Miltiadou, N. Koutras, D. Katsaros, P. Leškovský, L. Perlepes, A. Kostaridis, and F. Kouretas, "All-in-One Next-Generation Community Policing Solution Powered by Crowdsourcing, Data Analytics, and Decision Support: The INSPEC2T Case". in Social Media Strategy in Policing, Springer, Cham, 2019, pp. 217-251.
4. E. Charalambous, D. Kavallieros, B. Brewster, G. Leventakis, N. Koutras, and G. Papalexandratos, "Combatting cybercrime and sexual exploitation of children: an open source toolkit," in Open Source Intelligence Investigation, Springer, pp. 233–249, 2016.
5. E. Charalambous, "Application and development of computational intelligence methods in the analysis of archaeological data", in The NARNIA Project: Integrating approaches to ancient material studies, V. Kassianidou and M. Dikomitou-Eliadou, Eds. NARNIA Project, pp. 219–231, 2014.

Published conference proceedings

1. O. Gkotsopoulou, E. Charalambous, K. Limniotis, P. Quinn, D. Kavallieros, G. Sargsyan, S. Shiaeles & N. Kolokotronis, (2019). Data Protection by Design for Cybersecurity Systems in a Smart Home Environment. arXiv preprint arXiv:1903.10778.
2. S. Tsekeridou, G. Leventakis, G. Kokkinis, E. Charalambous, S. Anson, and G. Sargsyan, "A Crowd-Sourced Intelligent Information Management and Decision Support System Enabling Diverse E-Government G2C2G Services," in Third International Congress on Information and Communication Technology, 2019, pp. 687–705.
3. G. Leventakis, G. Papalexandratos, G. Kokkinis, E. Charalambous, and N. Koutras, "Towards efficient law enforcement decision support systems in the area of community policing: The use of mobile applications," in 2016 European Intelligence and Security Informatics Conference (EISIC), 2016, p. 198.

4. E. Charalambous, R. Bratskas, G. Karkas, A. Anastasiades, and N. Koutras, "An innovative Digital Forensic Tool assisting evidence analysis in Cyprus," *Big Data, Knowl. Control Syst. Eng.*, p. 45, 2015.
5. E. Charalambous, N. Efstathiou, and N. Koutras, "A cost effective solution for audio surveillance using embedded devices as part of a cloud infrastructure," *Big Data, Knowl. Control Syst. Eng.*, p. 21, 2015.
6. E. Maltezos, M. Skitsas, E. Charalambous, N. Koutras, D. Bliziotis, & K. Themistocleous, "Critical infrastructure monitoring using UAV imagery", in *Fourth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2016)*, 2016, vol. 9688, p. 96880P.
7. C. Hadjistassou, R. Bratskas, N. Koutras, A. Kyriakides, E. Charalambous, & A. M. Hadjiantonis, "Safeguarding critical infrastructures from cyber attacks: A case study for offshore natural gas Assets," *J. Polish Saf. Reliab. Assoc.*, vol. 6, 2015.
8. E. Charalambous, J. Takaku, P. Michalis, I. Dowman, & V. Charalampopoulou, "Automated motion detection from space in sea surveillance," in *Third International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2015)*, 2015, vol. 9535, p. 95350J.

Unpublished

1. L. Charalambous, "Digging Deeper into Data Processing with Emphasis on Compositional and Microstructure Data: Machine Learning in Support of Archaeological Analysis."
2. E. Charalambous, D. Eliades, and G. Mitsis, "An Introduction to Archaeological Data Analysis with Emphasis on Ceramic Compositional Data."
3. E. Charalambous, G. Mitsis, and D. Eliades, "A review of cluster analysis algorithms for archaeological ceramic data."

Elisavet Charalambous

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Hypothesis & Contribution	3
1.3	Thesis Outline	5
2	Classification in Supervised Learning	7
2.1	Learning Approaches	7
2.1.1	Supervised Learning	8
2.1.2	Unsupervised Learning	9
2.2	Significance Test	10
2.3	Distance Metrics & Validity Indices	11
2.4	Error Rate & Generalisation Loss	13
2.5	Assumptions & Overlooks in Classification	14
2.6	Data Uncertainty & Outliers	15
2.7	Summary	16
3	Classification Algorithms & State of the Art Analysis	17
3.1	Introduction	17
3.2	The Evolution of Classification Methods	18
3.3	Classification Evaluation	21
3.4	Distance Metrics	23
3.5	Feature Selection	26
3.6	Modeling Data Uncertainty	30
3.7	Classification Robustness	32
3.8	Statistical Hypothesis Testing	35
3.9	Summary	38

4	A Statistically Unbiased Classification Methodology for Robust Classification	41
4.1	Dataset Characteristics	43
4.2	Preliminary Design Process	44
4.2.1	Clustering Algorithms	45
4.2.2	Experimental Results	48
4.2.3	Key Findings	54
4.3	A Methodology for Robust Classification	55
4.3.1	The Dataset	56
4.3.2	Significance Testing with Simulation	56
4.3.3	Algorithm Parametrisation	59
4.3.4	Classification Evaluation	59
4.3.5	Bootstrapped 5×2 cv t-test and F-test	61
4.3.6	Estimation of t and F statistics for significance testing	62
4.3.7	Feature Selection	64
4.4	Identification of inter-class relationships	65
4.5	Summary	65
5	Case study I: Analysis of Compositional Archaeological Data with Uncertainties	69
5.1	Introduction	69
5.2	Compositional Archaeological Data under Uncertainty	71
5.2.1	Chemical Compositional Data	72
5.3	Classification in Archaeology	74
5.3.1	Problem Formulation	75
5.3.2	Analysis Practices for Compositional Data	75
5.3.3	Data Analysis in the Simplex	76
5.4	Validation of Methodology on Ceramic Data	78
5.4.1	The Archaeological Dataset	79
5.4.2	Hypothesis Testing	81
5.4.3	The experimental design	81
5.4.4	Statistical Testing	85
5.4.5	Results and Discussion	85
6	Case study II: Acoustic Event Detection	91
6.1	Introduction	91

6.2	Audio Coding	93
6.3	Classification in Acoustic Event Detection	94
6.3.1	Problem Formulation	94
6.3.2	Practices in Acoustic Event Detection	95
6.3.3	Feature Extraction for AED	98
6.3.4	Acoustic Event Detection for Surveillance	99
6.4	Validation of Methodology on Audio for Acoustic Event Detection	101
6.4.1	Implementation of robust audio analytics for surveillance	101
6.4.2	System Operational Information	102
6.4.3	Feature Extraction	104
6.4.4	Classifier Training Methodology	104
6.4.5	Classification Algorithms used in the experiment	105
6.4.6	Lightweight Audio Analytics	107
6.4.7	Cloud side analysis	109
6.5	Experiment Results & Conclusions	111
7	Data Analysis Suite Tool	113
7.1	Introduction	113
7.2	Motivation	114
7.3	Tool Design	114
7.4	Analysis Principles	116
7.4.1	Data Source Processing	117
7.4.2	Exploratory Data Analysis	118
7.4.3	Model Training & Analysis for Classification Problems	120
7.5	Future Developments & Application	127
8	Conclusions & Impact	129
8.1	Future Work	132
	Bibliography	135

Elisavet Charalambous

List of Figures

4.1	The unordered and ordered D matrix: the input and output of VAT . . .	48
4.2	K-means: External Validity Indices	52
4.3	FCM: External Validity Indices	52
4.4	KFCM Polynomial Kernel: External Validity Indices	53
4.5	KFCM RBF Kernel: External Validity Indices	53
4.6	The effectiveness of CLODD for different datasets	54
4.7	Flowchart of methodology	57
4.8	AUC - ROC Curve [184]	60
5.1	Spectrum of Analysed Artefact ©Maria Dikomitou-Eliadou	73
5.2	Microstructure of Artefact ©Maria Dikomitou-Eliadou	73
5.3	Representation of data in the constrained simplex space	77
5.4	Decision Tree example based on the values of three chemical elements .	83
5.5	Variability in classification accuracy between algorithms	86
5.6	Two dimensional plot of the sample generated with LDA. The fabrics are not distinct in the state space, many classes overlap	87
6.1	A sampled and quantised sound wave	93
6.2	Lightweight analytics on ES	103
6.3	Embedded System Audio Analytics Functional View	107
6.4	Lightweight analytics on ES	109
7.1	Django Architecture diagram.	115
7.2	Conceptual Design	115
7.3	Analysis Domain	116
7.4	Data Input Panel	117
7.5	Imported dataset detail view	118
7.6	Detailed view of clustering result	119

7.7 Snippet of training panel parametrisation 121

7.8 Model training through the data analysis suite. 122

7.9 Detailed page of training request 125

7.10 Classification of unseen data 126

7.11 Data analysis flow through the data analysis suite. 127

7.12 Data analysis flow through the data analysis suite. 127

7.13 Snippet of results for a file 128

Elisavet Charalambous

List of Tables

3.1	Approaches to robust classification	39
4.1	Estimate Iterations	58
5.1	The estimated accuracy and Jaccard index of each algorithm. The scores represent the mean score of all iterations.	88
5.2	Inter-class relationships in a multi-class problem.	89
5.3	Classification accuracy and Jaccard index scores when classification is performed on elements with mean concentration $>0.1\%$ and $<0.1\%$. . .	90
6.1	Top 3 configurations obtained in an experiment which involved the parameterisation of the audio analytics module.	108
6.2	Classification performance for the detection of events.	111

Elisavet Charalambous

Chapter 1

Introduction

The human brain is the most powerful pattern recognition machine, that it is currently impossible to achieve similar performance with computational methods. Machine learning (ML) is defined as a set of methods that can automatically detect patterns in data and subsequently utilise this knowledge to either predict new data or make decisions under uncertainty [180]. Pattern recognition is a branch of ML that focuses on the recognition of patterns and regularities in data [29]. Pattern recognition/ classification is the scientific field that concerns the development and implementation of computational algorithms which achieve pattern classification (how do we categorise data based on a number of attributes), feature identification and generation (identify the most informative data characteristics) and regression (quantitative describe possible interrelations between features).

ML algorithms have many applications – due to its classification and clustering capabilities – and are being deployed in interesting ways to either predict new data or make decisions under uncertainty. It has become increasingly ubiquitous with more and more applications even in the most unlikely places. A few simple applications of ML with high impact are anomaly detection, automated categorisation and trend revealing [95]. ML assists in anomaly detection to flag any malpractice even in very high volume high frequency data transactions/ communications as ML powered systems can detect a possible insider trading in a stock market, while it may also flag a rogue customer transaction as a fraudulent transaction in high volume business doing market place websites. Classification methods on the other hand neatly segregate under topics thousands of sources of news articles aggregating portals while marketing companies use ML to group customers into segments. The list of possible applications

is endless nowadays touching many and versatile areas: speech recognition, genetics, signal denoising, weather forecasting, image processing, face detection/ recognition, autonomous systems, or even the automobile industry.

Algorithms in the field are mainly discriminated into two categories: unsupervised and supervised learning each imposing a number of assumptions with however fuzzy boundaries leaving room for methods which combine aspects of both practices. Most research in the field is tailored made to the needs of its application domain. As a result, performance greatly degrades when solutions are applied in real environments, while domain experts struggle with the variety and complexity of their datasets. As far as this thesis is concerned when referring to supervised learning methods we refer to classification techniques, which is of our main interest, while when referring to unsupervised methods we refer to clustering approaches.

1.1 Motivation

The use of ML in so many domains and applications has led to an emerging need of application domain experts to, not only familiarise themselves with the basic principles of data analysis, but also to gain substantial comprehension on data sampling, transformation and analysis practices. The high demand in ML solutions leads to the emerging need for plug-n-play solutions that do not require extensive setup or re-training, whilst also allowing their integration in greater systems. The high performance expectations and, usually timely and data constrained, implementation strategy of ML solutions, lead researchers in not always following sound research procedures.

Additionally, it is often the case that data emerge as the result of observation or projection of artifacts – tangible or not – that exist (in physical form) or may be perceived in the multidimensional real world. The sampling procedure of the artifacts to their quantitative or qualitative representation generally introduces uncertainty. Usual sources of uncertainty are the sampling instrumentation used and range of selected parameters/features thought to be representative enough for the artifacts in question. These factors add to the uncertainty that the artifact may be imperfect or imprecise even in its original form. Considering, the complexity and involved factors, it is rather challenging to train robust classification models with validated performance without coding or scripting involved.

Robust classification of data with uncertainties is thought to be possible when

enough parameters of the application domain are considered and analysed. Throughout this thesis, and particularly in Chapter 3 multiple cases where various factors influence the classification outcome will be disclosed. The nature of data, pre-processing operations, feature aggregation and data transformations are factors that impact significantly the results of analysis. Even though solutions based on supervised learning are widely deployed, it is not necessarily implied that the above mentioned aspects were thoroughly analysed or that classifier learning followed a sound procedure.

As no classification algorithm is suitable for all problems and based on the fact that increased model complexity does not necessarily imply improvements in performance, there is an emerging need for methods that deal with the training of robust classifiers. Moreover, the era of big data and high metadata fidelity, leads to an abundance of available features that may constitute a sample (highly dimensional and heterogeneous feature spaces) and uncertainty in the methods used for recording. As a result, situations with heterogeneity in the nature of data and feature vectors are rather common.

From a practical point of view, and as far as this thesis is concerned, robust classification involves an algorithm's ability in tolerating small input changes that might have been caused either due to sampling uncertainties or due to the unavailability of enough data. Even though various robust optimisation techniques exist in the literature, these are not considered in this thesis as they are algorithm specific and therefore could not be employed in the context of an algorithm agnostic methodology. However, the employment of robust optimisation when an appropriate algorithm is selected, for learning with the proposed methodology, is still valid and possible. Additionally, this thesis is also concerned with the robustness of the training outcome produced as a result of utilising well established, and sound steps to alleviate the likelihood aspect introduced by an algorithm's initial and training-test set conditions during learning.

1.2 Hypothesis & Contribution

Based on the challenges and motivating factors discussed in Section 1.1, this thesis contributes to the research community by proposing a structured and systematic – algorithmic agnostic – methodology that deploys sound methods for the training of classifiers with evaluated robustness. The proposed methodology allows robust learning – in the sense that it alleviates the influence of likelihood – of one or more classifiers with algorithms of preference and their performance evaluation. The above-mentioned

aspects present novelties in the area of machine learning.

An additional aspect of the suggested methodology is that it is statistically unbiased, in the sense that all deployed algorithms are treated equally. Data re-sampling in each iteration precedes the training and testing of each algorithm and therefore the same training, validation and testing data are provided. Additionally, flexibility is allowed to the fine-tuning and parametrisation of each involved classification algorithm based on their intrinsic characteristics.

Additionally, pair-wise comparative analysis of the classification outcome to determine significance in the output is also possible for the selection of the best performing classifier. Designation of this methodology followed investigation and analysis of factors and parameters that influence the classification result, including dataset characteristics and inter-class relationships.

The systematic learning approach also allows a researcher to evaluate an annotated dataset's ability to sufficiently discriminate between the different categories. Doing so, should allow the researcher to evaluate the expert's labeling – which usually considers a number of additional attributes – solely by the underlying structure of the dataset.

The novelty of the suggested methodology lies on its statistically valid and unbiased design which considers the idiosyncrasies of scarce heterogeneous data whilst also acting as a validated model, for its robustness, to allow the reliable categorization of new samples; to the nearest class or classes.

Having developed such a methodology, it is subsequently of interest to validate its applicability to multiple domains, without any additional steps or transformations, and to validate its effectiveness with the use of different classification algorithms on the same data.

The proposed methodology is applied in the fields of archaeology and security through two independent case studies that share data uncertainty and sparsity as common characteristics. The first case study is focused on the analysis of compositional archaeological data with uncertainties while the second is focused on the analysis of audio waveforms for the detection of key events in the area of acoustic event detection (AED) for surveillance.

The effectiveness of the suggested methodology will be tested against the null hypothesis that it is possible to test robustness of classification of data with uncertainties in an algorithmic agnostic approach, while the alternative hypothesis states that domain-specific approaches need to be developed in order to produce robust classifiers.

1.3 Thesis Outline

This thesis is organised as follows. Chapter 1 covers the introduction to the subject as well as the motivation and contribution of this study in the research community. Chapter 2 provides background on supervised learning with emphasis on the classification problem. Chapter 3, then provides an overview of current methods and best practices that aim to surpass limitations of existing approaches. Having defined the problem and investigated current practices, Chapter 4 introduces a statistically unbiased classification methodology implemented for robust analysis in an algorithmic agnostic approach. It additionally, discusses how the designation of an ensemble classification methodology may serve towards a more knowledgeable inference of the classification result. Finally, the methodology for performing robustness evaluation on classification for heterogeneous scarce data with uncertainties is presented. The application of the proposed methodology is discussed in Chapters 5 and 6 through two case studies. Case Study I is focused on the analysis of ceramic archaeological data, while Case Study II is focused on the analysis of sound waves for acoustic event detection in surveillance systems. The characteristics and requirements in each application domain are discussed and an appropriate methodology configuration is suggested.

The positive impact and successful validation of the methodology led to the designation of the Data Analysis Suite web-based tool, presented in Chapter 7 that allows application experts to perform exploratory analysis and train robust classification learners for binary problems. Finally, in Chapter 8 conclusions are drawn and the impact of this work is discussed.

Elisavet Charalambous

Chapter 2

Classification in Supervised Learning

The problem of searching for patterns in data is a fundamental one and has a long successful history [29]. An agent is learning if it improves its performance on future tasks after making observations about the world. In this chapter, the learning problem to train a function, given a set of input - output pairs, to predict the output for new inputs will be analysed. Even though this problem seems restricted, it exhibits vast applicability. Any component that adheres to intelligent characteristics, may be improved by learning from data. The characteristics of the component, the existence of prior knowledge, the representation used for the data and component as well as the availability of analysis feedback are contributing factors to the selection of the appropriate learning method.

2.1 Learning Approaches

Machine learning is usually divided into three main types, the predictive/ supervised learning approach – which is of main interest in this thesis –, the descriptive/ unsupervised learning approach and reinforcement learning. Additionally, class memberships range from crisp labels (which can be seen as a strong supervised learning setting) to the uniform class membership distribution $y_j = \frac{1}{L}$ for all possible classes (which can be considered as a special unsupervised scenario), while learning with uncertain class labels, or with weak teaching signals can be seen as a special type of partially supervised learning (PSL) [227].

Even though, supervised learning is the focus of this thesis, unsupervised learning methods through clustering and other exploratory analysis techniques are key and allow insight in dataset characteristics. It is common to train and apply, multiple such

techniques on the under analysis dataset, to observe characteristics of relevant to data distribution, feature dominance, separability and cluster dispersion. Such information is critical to the appropriate selection of the best classification algorithm for the specific problem; considering that the dataset is representative enough.

2.1.1 Supervised Learning

In supervised learning, the goal is to learn a mapping from inputs x to outputs t , given a labeled set of input-output pairs $D = (x_i, t_i)_{i=1}^N$ with D being the training set and N the number of training examples. Given a training set of N example input-output pairs

$$(x_1, t_1), (x_2, t_2), \dots (x_N, t_N)$$

where each t_i was generated by an unknown function $t = f(x)$, discover a function h that approximates the true function f . In this context, x and t can be any value, the function h is a hypothesis. Learning is a search through the space of possible hypotheses for one that will perform well, even on new examples beyond the training set. Over-training a learner through multiple iterations over the training set may lead to over-fitting; a phenomenon that occurs with all types of learners, even when the target function is not at all random. Overfitting becomes more likely as the hypothesis space and the number of input attributes grows and less likely as the number of training examples is increased [214].

The development of robust pattern classifiers from a limited training set $T = \{x_1, \dots, x_m\}$ of observations (i.e., feature vectors) $x_i \in X$, represented in a proper feature space X , has long been one of the most relevant and challenging tasks in machine learning and statistical pattern recognition [132]. A successful supervised learning algorithm is expected to accurately predict the target class for any data vector x . Each training input x_i is a p -dimensional vector, in general however this could be a complex structured object such as an image, a time series, a molecular shape or in the case of archaeology the chemical composition of a specimen.

In the supervised framework, any given generic observation $x \in T$ is uniquely associated with a corresponding target label $y \in Y$. It is assumed that X is a real-valued vector space such that $X \subseteq \mathbb{R}^p$, and that $Y = \{y_1, \dots, y_L\}$ is the set of L different class labels reflecting the ground truth of the classification problem at hand. The output variable y_i can in principle be anything, however most methods assume that y_i is a categorical or nominal variable from some finite set in the case of classification

$y_i \in \{1, \dots, Y\}$ (i.e. fabric A or fabric B) or a real valued scalar (i.e. exposure level, specimen age or risk level) in the case of regression; regression is beyond the scope of this thesis. Intervention from human experts is needed in order to annotate the training set correctly.

In classification, the training set is fed into a pre-selected supervised learning algorithm aimed at training a classifier C , that is a mapping $C : \mathbb{R}^p \Rightarrow Y$. This algorithm is expected to exploit the information encapsulated within both the feature vectors and the corresponding class labels [227]. Besides the training algorithm, a hypothesis space has to be fixed, as well. The hypothesis space consists of all the potential candidate classifiers C which may be the eventual outcome of the computation of the learning algorithm on the training set [9]. A hypothesis generalises well if it correctly predicts the value of y for novel examples. Sometimes the function f is stochastic in the sense that it is not strictly a function of x , and what is required is to learn the conditional probability distribution $P(Y|x)$.

2.1.2 Unsupervised Learning

On the other hand, an operational definition unsupervised learning can be stated as follows: given a presentation of n inputs, find c groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between artifacts (samples) in different groups are low. Considering a set of n sample measurements $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^p$, where the coordinates of x_i provide feature values. We assume that there are groups (subsets) of similar samples in X (the clusters) which however do not bear any class identifier.

The process of discriminating unlabeled data seeks solution to two problems. The first, is concerned primarily on the clustering approach and involves: assessing cluster tendency, partitioning and cluster validity. In other words, one should first determine the number of clusters present, then determine which objects belong to each one, and to what degree, and finally validate how good is the partitioning. Assessing cluster validity is of great importance and the performance of clustering methods greatly depends on specifying the parameters correctly. The second problem is concerned purely with the way similarity between the different samples in X is measured. An ideal cluster can be defined as a set of points that is compact and isolated [130].

Possible solutions to the clustering problem requires an integer number c repre-

senting the number of clusters which can be either crisp or fuzzy partitions. Crisp clustering can be formulated, in general, as a problem of partitioning the finite set X into a given number c of disjoint clusters. The crisp c -partitions of X are sets of cn (remember c is the number of clusters and n is the number of analyzed artifacts in set X) membership values u_{ik} . This results in a $c \times n$ membership matrix in which element u_{ik} defines the membership of sample x_k in cluster i . The set of all non-degenerate (no zero rows) c -partition matrices for X is:

$$M_{hc} = \{U \in \mathfrak{R}^{cn} | u_{ij} \in \{0, 1\} \forall i, j; \sum_{i=1}^c u_{ij} = 1 \forall j; \sum_{j=1}^n u_{ij} > 0 \forall i\} \quad (2.1)$$

where the partition element $U_{ik} = 1$ if x_k is labelled i and is 0 otherwise.

In unsupervised learning, the only provided input is set $X = \{x_i | x_i \in \mathbb{R}^p, i = 1, \dots, n\}$ and the goal is to discover underlying structures in the data. This is a less well-defined problem as no target groups or indications are being provided. Unsupervised learning also imposes challenges in suggesting suitable evaluation metrics as the comparison of the prediction of y for a given x to the observed value is not possible. Reinforcement learning is useful for learning how to act or behave when given occasional reward or punishment signals [18]; in the form of a scalar reinforcement signal that constitutes a measure of how well the system operates. The learner is not told which actions to take, but rather must discover which actions yield the best reward, by trying each action in turn [147].

2.2 Significance Test

Statistical significance testing is used to test whether the information gain between training iterations or methods is statistically unlikely. Statistical hypothesis testing methods allow the inference of a hypothesis ensuring that the predicted result is unlikely to have occurred by chance alone, according to a predetermined threshold probability [58]. Statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn.

Such a test begins by assuming that there is no underlying pattern (the so called null hypothesis). The data are analysed to calculate the extent to which they deviate from a perfect absence of pattern. If the degree of deviation is statistically unlikely (usually depicted by a probability of 5% or less), it is considered to be good evidence for the

presence of a significant pattern in the data. The probabilities are calculated from standard distributions of the amount of deviation one would expect to see in random sampling. In this context, the null hypothesis states that no information gain is to be observed even if an indefinite large sample was provided. The probability under the null hypothesis for a sample of $v = n + p$ would exhibit the observed deviation from the expected distribution of positive and negative samples. The deviation of comparing actual values of positive and negative examples in each subset p_k and n_k , with the expected numbers, \hat{p}_k and \hat{n}_k , assuming true irrelevance (see equation 2.2).

$$\hat{p}_k = p \times \frac{p_k + n_k}{p + n} \quad \hat{n}_k = n \times \frac{p_k + n_k}{p + n} \quad (2.2)$$

A convenient measure for the total deviation is given by equation 2.3.

$$\Delta = \sum_{k=1}^d \frac{(p_k - \hat{p}_k)^2}{\hat{p}_k} + \frac{(n_k - \hat{n}_k)^2}{\hat{n}_k} \quad (2.3)$$

Under the null hypothesis the value of Δ is distributed according to the χ^2 distribution with $v - 1$ degrees of freedom. The χ^2 table may be used to tell if a particular Δ value confirms or rejects the null hypothesis.

The methods of inference used to support or reject claims based on sample data are known as tests of significance. Statistical hypothesis testing is necessary in assuring, with a certain degree of confidence, that the outcome is not random, also allowing performance comparisons between different method.

2.3 Distance Metrics & Validity Indices

Many ML algorithms, heavily rely on the distance metric for the input data patterns due to the need that good metrics reflecting reasonably well the important relationships between the data need to be given [272]. Distance Metric learning is to learn a distance metric for the input space of data from a given collection of pair of similar/dissimilar points that preserves the distance relation among the training data. In recent years, many studies have demonstrated, both empirically and theoretically, that a learned metric can significantly improve the performance in classification, clustering and retrieval tasks [276].

Previous work [114,116,178] has shown that appropriately designed distance metrics can significantly benefit classification accuracy compared to the standard Euclidean

distance [276]. Extensive work has been undertaken in the development of distance learning methods aiming optimal results, however the majority of research focuses on unconstrained spaces where the analysis of compositional and frequency domain or transformed data is not appropriate leaving only a few possible options.

Depending on the availability of the training examples, algorithms for distance metric learning can be divided into two categories: supervised distance metric learning and unsupervised distance metric learning. The training examples of supervised distance metric learning is cast into pairwise constraints: the equivalence constraints where pairs of data points that belong to the same classes, and in-equivalence constraints where pairs of data points belong to different classes whereas for unsupervised distance metric learning, the idea is to learn an underlying low-dimensional manifold where geometric relationships (e.g. distance) between most of the observed data are preserved. There is deep connection between unsupervised distance metric learning and dimension reduction [276].

Since, classification (supervised) is a more constrained problem than clustering (unsupervised), clustering techniques may still be applied to classification problems to draw information related to the distribution of clusters. Key information relevant to the obtained clusters or classes, depending on the nature of the problem, may be obtained by combining compactness and separability metrics. Compactness measures the closeness of cluster elements and it is usually measured by variance. Separability indicates how distinct two clusters are, measuring the distance between representative objects of two clusters. Such aspects are of concern to research around clustering validity indexes [206].

There are three approaches to study cluster validity [249]. The first is based on external criteria of a pre-specified structure imposed on a dataset (external information that is not contained in the dataset). The second approach is based on internal criteria with the use of information that involves dataset vectors. Internal criteria are subdivided into two groups: the one that assesses the fit between the data and the expected structure and others that focus on the stability of the solution [188]. The third approach of clustering validity is based on relative criteria, which consists of evaluating the results (clustering structure) by comparing them with other clustering schemes.

2.4 Error Rate & Generalisation Loss

Model training involves the iterative process of finding the best fit of a model to the training dataset. In practice, the model's performance is evaluated against different parametrisation. The error rate of a hypothesis is defined as the proportion of times that $h(x) \neq t$ for an input-output pair (x, t) . The error rate combined with the cost of error for each class (aka utility) constitute the loss function $L(x, t, \hat{t})$ defined as the utility of predicting $h(x) = \hat{t}$ when the correct answer is $f(x) = t$. However, the hypothesis h with the smaller error does not also imply good generalisation ability. A hypothesis is thought to generalise well when it consistently performs well on unseen data. This process usually involves the use of cross-validation methods, where the available data are randomly split into a training set from which the learning algorithm produces h and a test set on which the accuracy of h is evaluated [214].

The generalisation loss for a hypothesis h with respect to the loss function L is defined as equation 2.4, given that the prior probability distribution over all input-output pairs is defined as $P(X, Y)$ and $P(x, y)$ being the probability of vector x belonging to label y .

$$GenLoss_L(h) = \sum_{(x,y) \in D} L(y, h(x))P(x, y) \quad (2.4)$$

and the best hypothesis h^* , is the one with the minimum expected generalisation loss, equation 2.5

$$h^* = \arg \min_{h \in H} GenLoss_L(h) \quad (2.5)$$

However, because $P(x, y)$ is not known, the learning agent can only estimate generalisation loss empirically as:

$$EmpLoss_{L,D}(h) = \frac{1}{N} \sum_{(x,y) \in D} L(y, h(x)) \quad (2.6)$$

with the best hypothesis \hat{h}^* as

$$\hat{h}^* = \arg \min_{h \in H} EmpLoss_{L,D}(h) \quad (2.7)$$

Factors such as unrealizability, variance noise and computational complexity constitute the reasons why \hat{h}^* may differ from the true function f .

2.5 Assumptions & Overlooks in Classification

Usually machine learning methods operate under a set of basic assumptions to prove their effectiveness and validity. Discrepancies are however noticed between assumptions imposed during method designation and realistic conditions. Performance is usually evaluated with respect to the ability to reproduce known knowledge, while in Knowledge Discovery and Data Mining (KDD) the key task is the discovery of previously unknown knowledge [268]. Therefore, the most fundamental aspect in producing plausible classification is that representative class labels and sufficient data, for each category, need to be provided (also related to size of the feature space); the training sample needs to be representative to the data population. One of the most confusing things about understanding learning theory is the vast array of differing assumptions. There is a gap between the assumptions made to prove that methods work and the assumptions that are realistic in practice. However, due to the ubiquity of high dimensional problems, the gap has become dangerously wide. [265].

Assumptions such as data smoothness, Independent and Identically Distributed (IID) sequences as well as class membership are common in classification algorithms. Smoothness assumptions on the data are linked to the cluster assumption stating that data points in the same cluster are likely to belong to the same class [190] and the manifold assumption stating that a high-dimensional dataset can be embedded into a lower dimensional manifold [45, 234]. The smoothness of a classifier revolves around the notion that if two input data points (x_1, x_2) are close to each other, then the corresponding classifier outputs (t_1, t_2) are mutually close, as well.

Another very common assumption is that objects $x \in X$ are independently drawn from some (unknown, yet identical) probability distribution defined on $P(x)$ (i.i.d. assumption) [29]. The labels $t \in T$ are given for each object according to some (also unknown but fixed) function $\eta(x)$ also covering the more general situation where each object can have more than one possible label ($P(t | x)$) [257]. However, it turns out that many results of pattern recognition theory carry over a weaker assumption. Namely, under the assumption of conditional independence and identical distribution of objects, while the only assumption on the distribution of labels is that the rate of occurrence of each label should be above some positive threshold [216].

Another assumption linked with the fact that training data are representative to their distribution and characteristics, there is also the stationarity assumption that the

probability distribution of data remains stationary over time [214].

Finally, a fundamental assumption adopted by traditional supervised learning is that each example belongs to only one concept, i.e. having unique semantic meaning [278]. However as this is a very strong, and in most cases non-realistic, a number of methods have been developed over the years to account for the multiple semantic meanings that one real-world object might have; also escaping from the notion of partial (fuzzy) membership. The multi-label learning paradigm emerges by assigning a set of proper labels to the object to explicitly express its semantics [252]; a very useful approach for the classification of archaeological artefacts.

Deployment of classification and in general ML methods requires compliance to inherent assumptions and also consideration of underlying mechanics. Many learning problems are formulated as minimization of some loss function on a training set of examples, while loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances. The difference between the two arise from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples [153]. Therefore, deployment of a classification algorithm may serve as a solution only when its objectives are well tied to the objectives of the problem.

2.6 Data Uncertainty & Outliers

As one of the major issues in data mining, outlier detection, or anomaly detection, has found numerous applications in a variety of fields [263]. Outliers are the observations, events, or items which do not conform to an expected pattern or deviate from the majority of the data. They arise often due to human error, systematic changes, fraudulent behavior, or natural deviations in populations [50].

In recent years, many new techniques have been developed for mining and managing uncertain data. This is because of the new ways of collecting data which has resulted in enormous amounts of inconsistent or missing data. Such data is often remodeled in the form of uncertain data. The outlier detection problem is particularly challenging for the uncertain case, because the outlier-like behaviour of a data point may be a result of the uncertainty added to the data point. Furthermore, the uncertainty added to the other data points may skew the overall data distribution in such a way that true outliers may be masked [4].

However, in cases of scarce data, practices implementing outlier detection and removal may hinder the data mining outcome. When dealing with scarce datasets, one should not expect that classes are equally represented. Samples from such classes may be mistakenly detected as outliers. Their removal leads to significantly skewing the class distribution or even discarding the whole class. It is therefore of interest to utilise re-sampling methods to minimise the effect of possible outliers, without the discarding of samples.

2.7 Summary

In this chapter, the fundamental and key principles of learning with emphasis on supervised learning and classification in particular were presented. The supervised and unsupervised learning problems were defined and the concept of statistical significance testing as a method of ensuring that the predicted result is unlikely to have occurred by chance alone was explained. The often overlooked role of distance metrics in reasonably reflecting the important relationships in the data is also discussed as it will be further utilised in Chapter 4. Additionally, the estimation of error rate and generalisation loss were also introduced as a means of stating the importance of estimating a good hypothesis.

The introduction of concepts in this chapter is important to set the foundations for the rest of this document, as Chapter 3 will discuss methods and approaches used in principle in the area of classification and Chapter 4 will build on these concepts to introduce a learning methodology for robust classification in order to tackle common overlooks by researchers when dealing with classification problems.

Chapter 3

Classification Algorithms & State of the Art Analysis

3.1 Introduction

During the past decade, machine learning has been widely applied to diverse problems from automatic annotation of multimedia contents [33, 200, 218, 251, 264], to bioinformatics [55, 82], web mining [142], rule mining [247], information retrieval [101], tag recommendation [234], and many more. As over the years, new classification methods are introduced, it is important to create standardised workflows that allow the plausible and valid learning and deployment of classifiers. This idea is further reinforced by the fact that complex structures in the form of deep neural networks have become prevalent and are being widely deployed in real-world applications ranging from image classification [31, 49] to autonomous driving [151, 230].

The impact and effectiveness of ML methods becomes evident with their wide applicability in diverse problems. However, the robustness of an ML model, prior to its deployment, is a critical factor and the subject of investigative interest. Robustness was briefly introduced in Chapter 1 and refers to a model's ability in retaining consistency in terms of performance and immunity to small input changes which is also of interest in this work. Typical ways of measuring robustness in this context are the use of boosting and classification evaluation methods [63] in combination with significance testing [144] with the use of appropriate classification evaluation metrics. In Chapter 2, the various factors that influence the classification outcome have been analysed. This is further reinforced by the fact that, in real-world environments it is usually difficult

to specify target operating conditions precisely making building robust classification systems problematic since most of the times a realistic environment is neither isolated nor uncontrolled.

The practical application of classifiers in the pipeline of operations may impact the performance or judgement of other systems. Due to this, over the years, several methods related to re-sampling, classification fusion and multi-layered classification (cascaded and hierarchical) have been developed with the purpose to improve the performance and consistency of the trained classifiers. In the rest of this Chapter, the state-of-art analysis, with respect to relevant aspects, is presented with emphasis on aspects relevant to the proposed methodology which is discussed as part of the next chapter, Chapter 4.

3.2 The Evolution of Classification Methods

Over the years, a number of algorithms have been recognized, while in 2009 the following algorithms have been characterised as the top 10 most influential data mining methods in the research community [271]: C4.5, k-Means, SVM (Support Vector Machines), Apriori, EM (Expectation Maximization), PageRank, AdaBoost, k-NN (k-Nearest Neighbours), Naive Bayes, and CART (Classification And Regression Trees). ML approaches have been categorised under five distinctly different approaches: deterministic sensitivity analysis, probabilistic sensitivity analysis, Bayesian frameworks, fuzzy set theory, and grey theory [37]. Even though numerous classification methods exist and have been effectively used as solutions to diverse applications, their adaptation to serve the requirements of individual application domains was necessary. Such examples are endless, one is presented in [250] where a fuzzy-input fuzzy-output SVM (F2SVM) is introduced where fuzzy class memberships are used during the training phase, and a fuzzy output is generated by using a logistic transfer function. In [223] the F2SVM is applied in speech processing for classification of voice quality characteristics.

Additionally, recent work in unsupervised feature learning and deep learning has shown that being able to train large models can dramatically improve performance allowing the development of deep networks capable of performing training on billions of parameters using tens of thousands of CPU cores [64]. Deep artificial neural networks (including recurrent ones) have won numerous contests in pattern recognition and machine learning [225].

A standard neural network (NN) consists of many simple, connected processors each producing a sequence of real-valued activations. NN-like models have been around for many decades if not centuries, while models with several successive nonlinear layers of neurons date back to the 1960s and 1970s having as main landmark the appearance of the back propagation (BP) algorithm. Despite the potentials, BP-based training of Deep NNs with many layers, has been found to be difficult in practice and had become an explicit research subject. Deep Learning became practically feasible through the help of Unsupervised Learning and Deep NNs attract wide-spread attention since 2000, mainly by outperforming alternative machine learning methods such as kernel machines [258] in numerous important applications. Deep NNs have also become relevant for the more general field of Reinforcement Learning (RL). Deep learning is leading in many domains, due to its ability to achieve the accuracy of kernel SVMs with the scalability of stochastic gradient descent.

At the same time, different methods and algorithms have been proposed to train models in pattern recognition applications [13,106,222] using partially or weakly labeled training data sets. Many semi-supervised classification algorithms have been developed during the last 20 years, as a means of providing solutions to problems that could not entirely fit in any of the two major categories.

Due to the wide and diverse range of existing algorithms, there is emerging need for the implementation of frameworks, like the one proposed in this work, to allow the robust classification of classifiers with any preferred algorithm, regardless of their structure.

Innovations in the field of classifications do not only involve the invention or improvement of existing classification algorithms, but also innovative ways in the deployment of classifiers; these might involve classifiers deployed under a hierarchical or cascaded manner or classifier ensembles that utilise classification fusion or majority voting techniques in order to provide an output. Such efforts aid the improvement of accuracy in the final output by combining classifiers, however, they make even more significant the use of robust classifiers in such setups to avoid errors propagation.

Research in this area is divided into two broad groups, the first is the combination of classifiers that predict the same set of random variables while the second involves the incorporation of classifiers as components in large intelligent systems. The aim in the first group is to improve classifications by combining the outputs of the individual models. Boosting [93], in which many weak learners are combined into a highly accurate

classifier, is one of the most common and powerful such schemes. In contrast, the objective in the second group is to allow multiple classifiers to operate in harmony and smoothly within a unified environment. Kumar and Hebert in [154], present such an example where a large MRF(Markov Random Fields)-based probabilistic model was developed, linking multi-class segmentation and object detection.

In the machine learning community it is well known that more complex classification functions yield lower training errors yet run the risk of poor generalization. If the main consideration is test set error, structural risk minimization provides a formal mechanism for selecting a classifier with the right balance of complexity and training error [60]. In particular, the hierarchical structure was initially used to train different second-level classifiers. In the hierarchical case, a model is learned to distinguish a second-level category from other categories within the same top level. In the flat non-hierarchical case, a model distinguishes a second-level category from all other second-level categories [77]. Hierarchical decomposition of a classification problem allows for efficiencies in both learning and representation. Each sub-problem is smaller than the original problem, and it is sometimes possible to use a much smaller set of features for each sub-problem [146].

Additionally, the concept of Cascaded Classification Models (CCM) involves the employment of repeated instantiations of “black box” classifiers at each level for, a sub-problem or a set of sub-problems, combined to improve performance on some or all tasks. Specifically, the CCM framework creates multiple instantiations of each classifier, and organizes them into tiers where models in the first tier learn in isolation, processing the data to produce the best classifications given only the raw instance features. Lower tiers accept as input both the features from the data instance, as well as features computed from the output classifications of the models at the previous tier [118].

Finally, an ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of classifiers. The main discovery is that ensembles are often much more accurate than the individual classifiers that make them up. A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse [110]. An accurate classifier is one that has an error rate of better than random

guessing on new x values. Two classifiers are diverse if they make different errors on new data points [69].

3.3 Classification Evaluation

So far we have seen that a wide selection of classification methods have been developed over the years in an effort to serve as tools for the solution of copious and simultaneously diverse problems. However, the effectiveness of a classification algorithm is not only dependent on its nature or parameterisation, but it is also largely dependent on the application domain (i.e. account the number of correct predictions from all predictions, account the instances in which the classifier correctly did not assign an artifact to a class it does not belong, do not account for instances in which the algorithm correctly did not assign a sample to a specific class). Evaluation of the validity and the plausibility of classification results is not only necessary but critical, while it should follow appropriate parameterisation and use of metrics.

Therefore a number of classification evaluation metrics have been developed to account for both supervised and unsupervised methods as well as for the different types of classification/clustering (i.e. crisp, fuzzy, binary, multi-class, multi-label). A common approach for the evaluation of both supervised and unsupervised results is with the use of validity indices. One of the fundamental challenges of clustering is how to evaluate results, without auxiliary information; both training and testing data are needed for the assessment of an algorithm's generalization capability.

In order to determine the input parameters that lead to clusters that best fit a given dataset, we need reliable guidelines to evaluate the clusters [207]. Currently, cluster validity indices research has drawn attention as a means to give a solution [109]. Clustering validation is a technique to find a set of clusters that best fits natural partitions (number of clusters) without any class information. Cluster validity indices can be defined based on three different criteria: internal, relative and external [131]. Indices based on internal criteria assess the fit between the structure imposed by the clustering algorithm and the data using the data alone. Indices based on relative criteria compare multiple structures (generated by different algorithms, for example) and decide which of them is better in some sense and evaluate the result with respect to information intrinsic to the data alone. External indices measure the performance by matching cluster structure to the a priori information and evaluate the result with

respect to a pre-specified structure (i.e., the ground truth). Finally, the third approach of clustering validity is based on relative criteria, which consists of evaluating the results (clustering structure) by comparing them with other clustering schemes [207].

Internal cluster validity indices can be further divided into two groups where the former assesses the fit between the data and the expected structure and the latter focuses on the stability of the solution [189]. The notion of cluster stability [157] is appealing as an internal stability measure. Cluster stability is measured as the amount of variation in the clustering solution over different sub-samples drawn from the input data. Consequently, different types of indices are used to solve different types of problems and indices selection depends on the kind of available information. In general, clustering validity indices are usually defined by combining compactness and separability. Compactness is the measure known describing closeness of cluster elements. A common measure of compactness is variance, while the measure of separability indicates how distinct two clusters are; it essentially computes the distance between two different clusters.

For a classification application with discrete states the performance of a classifier is usually summarized by a confusion matrix. The elements of the confusion matrix determine the number of samples correctly (or incorrectly) classified. An evaluation metric then summarizes this confusion matrix into a value that can be used for comparing different classification techniques or different models for the classifier [88].

This choice of the evaluation metric is very important and application-dependent. A poorly defined metric may guide the model selection procedure to a far-from-optimal model or lead to erroneous conclusions when comparing the performances of two classifiers. Several evaluation metrics, including the classification accuracy, mis-classification costs, Kappa coefficient [57], the receiver operating characteristics (ROC) curve [284], and loss functions have been proposed. For evaluating classification problems with balanced datasets, classification accuracy is commonly used. Serious problems with classification accuracy arise when classes are highly imbalanced (e.g., for a two-class problem $p(C_1) \gg p(C_2)$) [199]. Imbalanced datasets are common and of interest to this thesis and therefore classification accuracy is not a suitable evaluation metric.

Unfortunately, choosing an alternative evaluation metric in classification applications with imbalanced datasets is not obvious. Each metric has strengths and weaknesses, however, research studies seldom justify why a particular metric was chosen for that specific application. Provision of an objective method for comparing metrics

would allow the informed selection of metrics suitable for the particular application.

In [125], a framework to compare (classification accuracy and AUC) two well-known metrics frequently used to summarize a confusion matrix is proposed. With AUC we refer to the area under the receiver operating characteristic (ROC) curve. This framework uses two measures: the degree of consistency (DoC) and the degree of difference (DoD) to compare classification accuracy and AUC. Based on this work, Fatourehchi et al. [88] proposed a general solution for comparing metrics used in binary classification applications with imbalanced datasets. The focus in this study was on metrics that summarise a single confusion matrix (compared to metrics such as AUC whose calculation is dependent on multiple confusion matrices).

The area under the ROC (Receiver Operating Characteristics) curve, or simply AUC, has been traditionally used in medical diagnosis and other sciences – ecology included – since the 1970s. It has recently been proposed as an alternative single-number measure for evaluating the predictive ability of learning algorithms. Despite the fact that for multiple years no formal arguments were given as to why AUC should be preferred over accuracy, [125] showed theoretically and empirically that AUC is a better measure (defined precisely) than accuracy. The significance of this study escapes this proof as it also demonstrated that the results of statistical tests of significance are impacted by the choice of the evaluation metric. In particular, it was demonstrated that Naive Bayes and decision trees that are very similar in predictive analysis do not adhere to the same results in AUC; Naive Bayes was significantly better than decision trees when AUC was used.

3.4 Distance Metrics

Many ML algorithms rely their operation in a form of similarity or dissimilarity metric among a set of items in a multidimensional space; either this being other data samples or a metric relevant to the distribution of the class/cluster. These measuring functions are known as distance metrics and allow the supervised or unsupervised learning algorithm to make data based decisions. A good distance metric helps in improving the performance of classification, clustering and information retrieval process significantly.

The use of the appropriate distance metric is extremely important as neither information present in the data should be ignored nor that effects of noise or outliers are exaggerated. Failing to select the appropriate distance metric, then unwanted features

in the data will have undue influence on the results, perhaps obscuring meaningful patterns [168].

Many ML algorithms use the Euclidean distance as the default distance function where the distance between vectors in Cartesian coordinates $p = [p_1, p_2, \dots, p_n]^T$ and $q = [q_1, q_2, \dots, q_n]^T$ are two points in the infinite Euclidean n-space, then the distance d is defined as:

$$d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (3.1)$$

The use of this distance metric is not appropriate neither in all problems nor with all ML learning methods.

The City-block distance (aka Manhattan distance) is a commonly used metric that lies in the taxicab geometry where the distance between two points is the sum of the absolute difference of the Cartesian coordinates. The taxicab distance d_1 , between two vectors, in an n-dimensional real vector space is defined as:

$$d_1(p, q) = \|p - q\|_1 = \sum_{i=1}^n |p_i - q_i| \quad (3.2)$$

Taxicab distance depends on the rotation of the coordinate system, but does not depend on its reflection about a coordinate axis or its translation and therefore is one of the preferred distance metrics used to assess the differences in discrete frequency distributions.

Proportion coefficients can also be used as distance metrics and represent coefficients expressed as proportions of the maximum distance possible. An instance of a proportion coefficient is the Jaccard index (aka Jaccard coefficient and Jaccard similarity) and may be thought of as the overlap between the area under curves. If A is the area under one curve, B is the area under the other, then the Jaccard coefficient, in set notation, is defined as:

$$JaccardIndex = \frac{A \cap B}{A \cup B} \quad (3.3)$$

Proportion coefficients as distance measures are foreign to classical statistics, which are based on squared Euclidean distances. Ecologists latched onto proportion coefficients for their simple, intuitive appeal despite their falling outside of mainstream statistics. Nevertheless, Roberts in [210], showed how proportion coefficients can be derived from the mathematics of fuzzy sets an increasingly important branch of mathematics.

Another, interesting to our field distance metric is the chi-square measure (χ^2), used in correspondence analysis. Chi-square histogram distance is one of the distance measures that can be used to find dissimilarity between two histograms and has been widely used in various applications such as image retrieval, texture and object classification, and shape classification [192]. If p and q represent the probability distributions of two events A and B with random variables, $i = 1, 2, \dots, n$, the chi-square measure between these two histograms is given by [192]:

$$\chi_{A,B}^2 = \frac{1}{2} \sum_{i=1}^n \frac{[p_i - q_i]^2}{p_i + q_i} \quad (3.4)$$

In histograms of many processes, the difference between large bins is less important than the difference between small bins and that should be reduced. The chi-square histograms take this into account [192]. The chi-square histogram distance comes from the chi-square statistics to test the fit between a distribution and observed frequencies.

At this point it is important to note that depending on the type of data – under analysis – and the performed operation/transformation, the most appropriate distance metric should be used. For instance if the similarity between images is to be measured based on their histograms, the chi-square distance metric should be used, while in case image coefficients are extracted and stored in a database, then distance metrics such as the Manhattan Distance and the Euclidean distance should be used; a similar case is explained by Fan & Wang in [87].

Another such example involves the use of Kullback–Leibler Divergence (KLD) to analyze the spectral structure of acoustic events for Acoustic Event Detection (AED). In [282] KLD based feature discriminative capability analysis was applied to understand the relevance of different feature components (in a speech feature set) for the AED task compared to speech recognition. The distance between the distributions associated with an acoustic event label and the other audio labels reveals the discriminative capability of the feature for that acoustic event. KLD, denoted by $D(p||q)$, is a measure (a “distance” in a heuristic sense) between two distributions, p and q , and is defined as the cross entropy between p and q minus the self entropy of p (see Equation 3.5).

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \quad (3.5)$$

KLD was used to measure the discriminative capability of each feature component for each acoustic event. Let $d_{ij} = D(p_{ij}||q_{ij})$ denote the divergence between the distribution of the i^{th} feature component for the j^{th} acoustic event and the global distribution

of the i^{th} feature component for all the audio. The global discriminating capability of the i^{th} feature component is defined by Equation 3.6 where P_j is the prior probability for the j^{th} acoustic event.

$$d_i = \sum_j P_j d_{ij} \quad (3.6)$$

Even though, many distance measures exist, it is important to know the domain of acceptable data values for each distance measure. Many distance measures are not compatible with negative numbers while other distance measures assume that the data are proportions varying between zero and one [168].

3.5 Feature Selection

In the past thirty years, the dimensionality of the data involved in machine learning and data mining tasks has increased explosively. Data with extremely high dimensionality has presented serious challenges to existing learning methods [113, 160]. With the presence of a large number of features, a learning model tends to overfit, resulting in performance degradation.

Dimensionality reduction is one of the most popular techniques to remove irrelevant and redundant features and are generally categorised into feature extraction and feature selection. According to whether the training set is labelled or not, feature selection algorithms can be categorised into supervised [233, 267], unsupervised [78, 175] and semi-supervised [275, 280].

Feature extraction approaches project features into a new feature space with lower dimensionality, where the newly constructed features are usually combinations of original features. Feature selection approaches aim to select a small subset of features that minimize redundancy and maximize relevance to the target (i.e select the subset of highly discriminant features). Common feature extraction techniques include Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA), while common feature selection techniques include Information Gain, Relief, Fisher Score and LASSO [241].

Both feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better generalised models, and decreasing required storage. Feature selection is superior in terms of better readability and interpretability compared to feature extraction techniques as feature selection

selects a subset of features from the original feature set without any transformation, and maintains the physical meanings of the original features allowing further analysis on the obtained features if desired.

Feature selection aims to select a subset of highly discriminant features; the relevance of features is assessed as the capability of distinguishing the given sample into the set of different classes. For example, a feature f_i is said to be relevant to a class c_j if f_i and c_j are highly correlated [241]. This procedure generally consists of four basic steps [161], namely, subset generation, subset evaluation, stopping criterion, and result validation. In the first step, a candidate feature subset will be chosen based on a given search strategy, which is then evaluated, in the second step, according to certain evaluation criteria. The subset that best fits the evaluation criterion will be chosen from all the candidates that have been evaluated after the stopping criteria are met. The chosen subset will finally be validated using domain knowledge or a validation set.

Supervised feature selection, which is of interest in this thesis, assesses the relevance of features guided by the label information but a good selector needs enough labeled data, which is time consuming; a requirement not always satisfied when data are characterised by huge dimensionality. Supervised feature selection methods mainly affect the training phase of classification and are broadly categorized into filter models, wrapper models and embedded models. The filter model separates feature selection from classifier learning so that the bias of a learning algorithm does not interact with the bias of a feature selection algorithm. It relies on measures of the general characteristics of the training data such as distance, consistency, dependency, information, and correlation. Relief [211], Fisher score [76] and Information Gain based methods [193] are among the most representative algorithms of the filter model.

The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features. In a wrapper method the classifier is used as a black box returning a feature ranking, therefore one can use any classifier which can provide the ranking of features. For practical reasons, a classifier used in this problem should be both computationally efficient and simple, possibly without user defined parameters [155]. If this requirement is not fulfilled, these methods may become prohibitively expensive to run for data with a large number of features. Finally, the embedded model performs feature selection in the learning time; model fitting and feature selection are performed simultaneously [44, 201]. This model was proposed to bridge the gap between the filter and wrapper models, as it incorporates

the statistical criteria to select several candidate feature subsets with a given cardinality and then select the subset with the highest classification accuracy [161]. Thus, the embedded model usually achieves both comparable accuracy to the wrapper and comparable efficiency to the filter model.

In real-world classification problem little to none prior knowledge is available with regards to class and class-conditional probabilities, while, little knowledge about relevant features exists. Due to this, many candidate features are introduced to better represent the domain, resulting in the existence of irrelevant/redundant features to the target concept. A relevant feature is neither irrelevant nor redundant to the target concept; an irrelevant feature is not directly associated with the target concept but affects the learning process, and a redundant feature does not add anything new to the target concept [136]. A good discussion outlining why finding all relevant attributes is important is given by Nilsson et al. in [185]. In many classification problems, it is difficult to learn good classifiers before removing these unwanted features due to the huge size of the data. Reducing the number of irrelevant/redundant features can drastically reduce the running time of the learning algorithms and yields a more general classifier. This helps in getting a better insight into the underlying concept of a real-world classification problem [241].

The all-relevant problem of feature selection which involves the identification of all attributes which are in some circumstances relevant for classification is more difficult than usual minimal-optimal one where the objective is to find a set of non-redundant features. One reason is that we cannot rely on the classification accuracy as the criterion for selecting the feature as important (or rejecting it as unimportant). The degradation of the classification accuracy, upon removal of the feature from the feature set, is sufficient to declare the feature important, but lack of this effect is not sufficient to declare it unimportant [155]. One therefore needs another criterion for declaring variables important or unimportant. Moreover, one cannot use filtering methods, because the lack of direct correlation between a given feature and the decision is not a proof that this feature is not important in conjunction with the other features [107]. One is therefore restricted to wrapper algorithms, which are computationally more demanding than filters. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset.

The feature selection phase might be independent of the learning algorithm, like filter models, or it may iteratively utilize the performance of the learning algorithms to

evaluate the quality of the selected features, like wrapper models. With the finally selected features, a classifier is induced for the prediction phase. Usually feature selection for classification attempts to select the minimally sized subset of features that does not significantly decrease the classification accuracy (or any other classification evaluation metric) whilst retaining as much as possible the original class distribution [241].

Ideally, feature selection methods search through the subsets of features and try to find the best one among the competing 2^p candidate subsets according to some evaluation functions [62]. However this procedure is exhaustive as it tries to find only the best one. It may be too costly and practically prohibitive, even for a medium-sized feature set size (m). Other methods based on heuristic or random search methods attempt to reduce computational complexity by compromising performance. These methods need a stopping criterion to prevent an exhaustive search of subsets.

Researchers in statistics [34, 75, 156, 171, 183] and pattern recognition [21, 65] have investigated the feature selection (aka feature subset selection) problem for decades, but most work has concentrated on feature selection using linear regression.

Sequential backward elimination, sometimes called sequential backward selection, was introduced in Marill & Green (1963). Kittler generalised the different variants including forward methods and stepwise methods. Branch and bound algorithms were introduced by Narendra & Fukunaga (1977). Finally, more recent papers attempt to use AI techniques such as beam search and bidirectional search [229], best first search [273] and genetic algorithms [255].

Many measures have been suggested to evaluate feature selection (as opposed to cross validation), such as adjusted mean square error, adjusted multiple correlation coefficient and the C_p statistic [166]. The search for the best feature subset can be improved by making assumptions on the evaluation function. The most common assumption is monotonicity, that increasing the subset can only increase the performance. Under such assumptions, the search space can be pruned by the use of dynamic programming and branch-and-bound techniques. The monotonicity assumption is not valid for many induction algorithms used in machine learning.

The terms weak and strong relevance are used to denote formulas that appear in one minimal derivation or in all minimal derivations. Moret [177] then defined redundant features and indispensable features for the discrete case. The definitions are similar to our notions of irrelevance and strong relevance, but do not coincide on some boundary cases. Determination was introduced in Russel [213, 215] under a probabilistic setting

and used in a deterministic non-noisy setting in Schlimmer [224] and may help analyse redundancies.

3.6 Modeling Data Uncertainty

The main challenge of uncertainty in design is that different types of uncertainty can occur at different stages of the design process. These different uncertainties can require distinctive handling and modeling techniques to understand their influence and importance [150]. Complex information systems are expected to have several types of uncertainties, such as fuzzy, probabilistic, and non-specificity [186, 187]. Fuzzy Uncertainty deals with the imprecision or vagueness associated with the occurrence of an event. In contrast, Probabilistic Uncertainty models the uncertainty of an event belonging to a crisp set. A model that integrates these two uncertainty types may be called Combined Uncertainty (named Total Uncertainty in [186, 187]).

To date, there has not been a principled way of modeling data uncertainty directly for classification problems in the literature. A hidden underlying assumption is that errors are confined to the output y , i.e., the input data are not corrupted with noise; or even when noise is present in the data, its effect is often ignored in the learning formulation [28]. Despite the previous statement, attempts in learning robust classifiers have been studied in an algorithm specific manner (i.e., the learning of robust SVM). However, for many applications, this assumption is unrealistic. Sampling errors, modeling errors and instrument errors may preclude the possibility of knowing the input data exactly.

A simple approach that will most likely be sufficient for most decisions is deterministic sensitivity analysis, although more complex approaches may be needed when multiple sources of uncertainty must be simultaneously considered [37]. Hence classification problems based on observed data in reality are expected to have noisy inputs. Due to this, many systems opt to provide estimates for the reliability of their outputs, which measure how uncertain each element of the outputs is. Such information is proven useful in the learning formulation problem and should be considered when the objective is to produce more accurate predictors.

Data uncertainty modeling assumes that data are corrupted by noise such that (x_i, y_i) with x_i being the corrupted input emerge from x'_i the un-corrupted data, independently of the output y_i so that (x'_i, y_i) is still a valid expression. Given that x'_i

follows distribution $p(x'_i, y_i|\theta)$, where θ is an unknown parameter estimated from the data. The objective is to model a distribution $p(x_i|\theta', \sigma_i, x'_i)$ where σ_i is the known estimate of uncertainty for x_i . The joint probability of (x_i, y_i) is obtained by integrating out the unobserved quantity x'_i :

$$p(x_i, y_i) = \int p(x'_i, y_i|\theta)p(x_i|\theta', \sigma_i, x'_i)dx'_i \quad (3.7)$$

The unknown parameters (θ, θ') can be estimated from the data using the maximum likelihood estimate which often leads to a very complicated formulation due to the integration over the unknown true input. Additionally, generalisation of this method to non-probability formulations is not straightforward and therefore other, more computationally efficient and tractable, methods should be applied; such methods involve approximations of equation 3.7. The penalisation effects of uncertainty modeling approaches leads to ignoring data that are thought to be very uncertain and perform the learning and prediction procedure on data less contaminated.

Over the years, multiple attempts have been proposed to tackle the problem of uncertainty modeling. The complexity of the uncertainty problem, led researchers in decoupling the problem into different areas for further investigation. For instance, [28] investigated a learning model based on support vector classification in which the input data is corrupted with noise. Authors, introduced a formulation of SVM with uncertain input based on the total least squares regression method. Empirical results showed that the newly formed method is superior to the standard SVM for problems with noisy input. On the same lines, but losing the assumption that t represents the ground truth, authors in [26] proposed a framework based on robust optimization methods to address classification problems whose data (both in features and in labels) are subject to error, the three most widely used classification methods: support vector machines, logistic regression, and decision trees. Robust optimization is a flexible framework for modeling uncertainty [23] and is arguably one of the fastest-growing areas of optimization in the last decade. Research is also focused in combined uncertainty methods, in particular, [11] discusses the use of combined uncertainty methods in the diagnosis of coronary artery disease using electrocardiogram (ECG) stress signals. Combined uncertainty computes a composite of two types of uncertainties, fuzzy and probabilistic.

Due to the great interest in the field and the increasing number of produced publications in the area, in 2015 a work has been published, reviewing attempts to deal with uncertainty in classification as part of Multi-Criteria decision analysis [37]. Based

on the definitions of uncertainty suggested by Briggs et al. [36] on the different types of uncertainty, studies in the area of uncertainty between 1960 and 2013 were reviewed and categorized based on the MCDA method used. Additionally, [150] in a work published in the International Conference on Engineering Design (ICED11), proposed a classification of the manifestation of uncertainty describing the different points of the design process with the aim to set a basis for shared understanding and characterization of uncertainty.

A plausible approach for dealing with noisy input is to use the standard learning formulation without modeling the underlying input uncertainty. This reasoning relies on the fact that the noise observed in both the training and testing data is equivalent and therefore will impact the learning process similarly; making it negligible.

Different approaches of uncertainty described in literature focus on different aspects and points of the design process and offer insights on different aspects [150]. While uncertainty modelling might be beneficial to well defined problems, enough data should be available for the proper estimation of the probability distribution functions in equation 3.7; it is also assumed that the distribution functions of both data and uncertainty do not change over time.

Moreover, deterministic uncertainty modeling approaches, that are more straightforward and most commonly used, rely on the assumption that labels t do not adhere to uncertainty and represent the ground truth; a very strong assumption for applications related to soft sciences and human perception. In some applications, it is not acceptable to neglect data even when these are thought to be highly uncertain; especially when dealing with scarce datasets and under-represented classes [228].

3.7 Classification Robustness

Noisy data and uncertainty impose great challenges in the learning of accurate ML models. Unfortunately, the commonality of uncertain data calls for the implementation of methods, whether these are algorithm agnostic or not, that allow for robustness in the operation of the ML model and as a consequence the operation of the overall system.

In most classification settings, the proportion of misclassified samples in the test set is the main performance metric used to evaluate classifiers. Classification robustness works study empirically and theoretically the robustness to different types of pertur-

bations, such as adversarial perturbations, additive random noise, structured transformations, or even universal perturbations. The prediction accuracy has been the long-lasting and sole standard for comparing the performance of classification models, however, recent studies have highlighted the lack of robustness in well-trained classifiers to adversarial examples [238] creating the emerging need to experiment with multiple robustness metrics, including the distortion and success rate. The robustness is usually measured as the sensitivity of the discrete classification function.

In real-world environments it is usually difficult to precisely specify target operating conditions making building robust classification systems problematic. In robustness the objective is to fit a model on contaminated data such that you find a fit as close as possible as the one you would have had without the outliers. Outlier detection tries to find all the outliers that matter in the sample. That is, all points that exert a disproportional pull on the fitted parameter of the model. Robustness solutions approach the problem from different perspectives, some aid at learning a classifier based on robust optimisation techniques for maximum separation (assuming the probability distribution of noise), while others aid at optimising the learning solution by combining multiple analysis approaches with the aim to make systems more tolerable to noise; contributions in this thesis belong to the second approach.

Many recent attempts exist in the literature with respect to classification robustness. In the area of artificial neural networks, the concept of combining multiple networks has been proposed as a new direction for the development of highly reliable neural network systems. The authors in [52] proposed a method for multi network combination based on the fuzzy integral. This technique non-linearly combined objective evidence, in the form of a fuzzy membership function, with subjective evaluation of the worth of the individual neural networks with respect to the decision. The experimental results with the recognition problem of on-line handwriting characters confirm the superiority of the presented method to the other voting techniques.

Provost in [199] proposes a way to build a hybrid classifier that promised to perform at least as well as the best available classifier for any target conditions; based on empirical evidence. The authors claim that in some cases the performance of the hybrid actually can surpass that of the best known classifier. The authors' proposed method is based on the comparison of classifier performance that is robust to imprecise class distributions and mis-classification costs. Provost's solution extends across a wide variety of comparison frameworks based on the ROC convex hull (ROCCH) method combin-

ing techniques from ROC analysis, decision analysis and computational geometry, and adapts them to the particulars of analyzing learned classifiers. The method minimizes the management of classifier performance data, and allows for visual comparisons and sensitivity analyses.

Viola & Jones in [262] proposed an approach based on a series of cascaded simple classifiers for extremely fast detection in domains where the distribution of positive and negative examples is highly skewed. Their subject of interest involved face detection and database retrieval applications. Their methodology included the use of AdaBoost as a mechanism for training each classifier in the cascade. Each trained classifier was designed to achieve high detection rates and modest false positive rates can yield a final detector with many desirable features: including high detection rates, very low false positive rates, and fast performance.

Ben-Tal in [22] approaches the problem from a different perspective and studies efficient methods for robust classification under uncertainty in Kernel Matrices through a study for the designation of SVM classifiers when the kernel matrix, K , is affected by uncertainty.

Motivated by the fact that real-world applications need to be resilient to arbitrary input data, an important line of work has developed the open-world learning framework that checks if the inputs are within the same distribution as training data (in-distribution examples), or if they come from a different distribution referred to as out-of-distribution examples [24, 25]. State-of-the-art open-world learning systems equip machine learning classifiers with out-of-distribution detectors, and an input example is processed for classification only if the input passes through those detectors. In recent years, the research community has developed several out-of-distribution detection mechanisms that are effective in distinguishing out-of-distribution inputs [119, 151, 158].

Sehwag et al. in [228] investigated evasion attacks in the open-world learning framework and defined out-of-distribution adversarial examples, which represent a new attack vector on machine learning models used in practice. Out-of-distribution learning frameworks aim to discard input examples which are not from the same distribution as the training data of machine learning classifiers. Through experiments, authors found that existing out-of-distribution detectors are insufficient to deal with this threat while they suggest further investigation on distance comparisons in feature space [135, 159] as part of their future work.

Finally, [81] proposed an optimisation approach for the robust classification of scarce

data in the area of face recognition for computer vision in the multi-sub space, using scarce representation techniques. Due to the fact that data for face recognition from multiple classes lie in multiple low-dimensional sub-spaces of a high-dimensional ambient space. Authors cast classification as a structured scarce recovery problem where the goal is to find a representation of a test example that uses the minimum number of blocks from the dictionary. The authors showed that transforming the face recognition problem to a structured scarce recovery problem can improve the results of the state-of-the-art face recognition algorithms, especially when we have relatively small number of training data for each class.

3.8 Statistical Hypothesis Testing

Statistical hypothesis testing methods allow the inference of a hypothesis ensuring that the predicted result is unlikely to have occurred by chance alone, according to a pre-determined threshold probability [58]. Statistical hypothesis testing is necessary in assuring, with a certain degree of confidence, that the outcome is not random, also allowing performance comparisons between different methods. Statistical inference allows analysts to assess evidence in favour or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance.

The inference procedure using dispersion metrics relies on statistical hypothesis testing, and therefore, on how well the null model represents neutral expectations. Currently, there exists an extensive number of null models that can be used to infer assembly processes, ranging from simple null models based on random shuffling of taxon labels [59, 102, 103, 143, 266], to dynamic null models [196] and analytical frameworks [236] that incorporate macroevolutionary processes such as speciation, dispersal, and extinction. However, even with more dynamic null models and simulation power, relying on statistical hypothesis testing and passing a significance threshold to infer an assembly process are problematic, in part due to the sensitivity between p-values and sample size and how we interpret “significance,” but also because each analysis of a particular data type and test statistic results in a measure of significance. Researchers are then responsible for integrating across a suit of hypothesis tests, some that may be significant while others are not, in order to draw an inference. Arguably, a model-based inference procedure is necessary to incorporate all data at once, rank models of com-

munity assembly by their relative support, and, importantly, incorporate uncertainty in model inference.

Statistical methods emphasizing formal hypothesis testing have dominated the analyses used by various disciplines to gain insight from data. One such discipline is ecology where the use of statistical hypothesis testing involves an integral component in their analysis. Variations of statistical hypothesis testing approaches have been developed over the years to be coupled with standardised approaches while it has been noted that journal editors nowadays require authors to quote the exact P values yield by their analysis, and let readers make their own interpretation [129].

Statistical hypothesis testing has gained great interest for several decades in multiple disciplines. The examples are numerous; for instance, [260] proposed a quantitative criterion for the termination of the estimation process for the Maximum Likelihood Estimator algorithm. Statistical hypothesis testing was used for determining the quality of the outcome. In particular, the authors in [239] proposed a new method of exploratory data analysis was developed based on the calculation of the AUC metric and non-parametric statistical hypotheses testing to detect statistically significant differences in the characteristics of the wave trains of the muscles' electrical activity. Another such example is presented by authors in [246] where authors discuss the use of statistical hypothesis testing to test phylogeographical hypotheses. Additionally, authors in [139], review the use of statistical hypothesis testing in biology and stress the fact the use of these methods is often emphasized disproportionately at the expense of the original goal of testing the experimental hypothesis.

Statistical testing can be performed in a number of ways depending on the nature of the problem, the availability of data as well as the ability to deploy re-sampling on the original dataset. The performance of a hypothesis test are characterised by Type I and Type II errors. A Type I error occurs if we reject the null hypothesis (in favor of the alternative hypothesis) when the null hypothesis is true. Type I error is denoted as $\alpha = P(\text{TypeIError})$ A Type II error occurs if we fail to reject the null hypothesis H_0 when the alternative hypothesis H_A is true. Type II error is denoted as $\beta = P(\text{TypeIIError})$.

In general, for every hypothesis test, it is desirable to:

- Minimize the probability of committing a Type I error. That, is minimize $\alpha = P(\text{TypeIError})$. Typically, a significance level of $\alpha \leq 0.10$ is desired.
- Maximize the power (at a value of the parameter under the alternative hypothesis

that is scientifically meaningful). Typically, with a desired value of power to be 0.80 or greater. Alternatively, minimisation of $\beta = P(\text{Type II Error})$, aiming for a type II error rate of 0.20 or less.

In the case of heterogeneous scarce data with a large number of features candidate solutions are the McNemar's test [169], k-fold cross validation [97] and the 5×2 cross validation paired $t - test$ [67]; a variation of the k-fold cross validation test. However the deployment of statistical testing does not remove the bias of the tested data. For this reason, bagging methods such as bootstrapping are used. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates [80].

Statistical hypothesis testing is a well known and established method for a few decades now. Due to its significance, it is still employed in many statistical and ML studies across multiple domains [3,12,30,108,128,253]. The abuse of statistical hypothesis testing in the case of multiple comparison procedures has already been the subject of a number of reviews in biological and medical journals, primarily in the case of mis-analysis of factorial designs and regression techniques [40,51,100,138,165,195]. There are, of course, a broad range of problems that yield to such methods, notably problems that are amenable to replicated, manipulative experiments. However, despite the success of these traditional approaches in analyzing data from designed experiments, there is an increasing appreciation among disciplines that a singular focus on manipulative experimentation and associated analyses compresses the range of questions that can address [121].

While extensive work also exists in reviewing alternatives to hypothesis testing including techniques for parameter estimation and model selection using likelihood and Bayesian techniques. These methods emphasize evaluation of weight of evidence for multiple hypotheses, multi-model inference, and use of prior information in analysis [121]. Such alternative involve the estimation and confidence intervals for determining the importance of factors, decision theory for guiding actions in the face of uncertainty, and Bayesian approaches to hypothesis testing and other statistical practices [137].

3.9 Summary

In this chapter, the state-of-the-art on the areas relevant to this thesis were discussed. Aspects relevant to the learning of classifiers and their deployment in real-environments were of particular interest. The evolution of machine learning techniques led to its wide applicability in diverse areas. As a consequence, new algorithms and analysis practices are continuously developed calling for the need to create standardised workflows that allow the plausible and valid learning and deployment of classifiers. Fuzzy algorithms were introduced over the traditional hard classification methods in order to cope with the fact that in real life categories or classes are not necessarily distinct or that a sample may rightfully belong to more than one labels. These methods usually emerge as alterations on the hard partitioning algorithm while variations of weakly or partially labels training sets also exist.

Research in the field of classification evaluation revealed that the classification accuracy, which is the most commonly used metric, is inappropriate for use in imbalanced datasets. The selection of the appropriate classification evaluation metric in learning impacts the performance of the classifier. Due to this, we have investigated other alternative metrics such as the Jaccard Index (appropriate in the case of fuzzy labels) and ROC AUC for greater separation in hard partitioning problems.

Uncertainty, either internal or external, in data impacts the performance of a learned model. To limit this impact, methods of either modeling data uncertainty or adjusting an algorithm exist in the literature. Once such example involves the employment of fuzzy classification with fuzzy input and output pairs.

The stable performance of a trained model in real environments is linked with its robustness. As it is usually the case for fuzzy algorithms, robust classifiers emerge as alterations to well proven classification methods where the learning of their parameters is optimised to increase robustness in small input changes. The complexity of the problem emerges from the fact that the conditions of a real-world environment are dynamic and difficult to estimate. As a result, the problem has been approached by different perspectives (see Table 3.1)

Despite the advances, there is still lack of standard methods that may allow the learning of multiple classification algorithms through an algorithmically agnostic learning procedure. Proposed methodologies should respect algorithm designation assumptions and should consider all available data samples to allow its application on scarce

Robust Classification Approaches		
Type	Description	Characteristic
Classification Optimisation Approaches	Optimisation of hyper-planes for maximum margin between classes	Its application is algorithm dependent
Hybrid Classification	Combination of trained classifiers for different classes	Tailor-made solution for each problem. The learning and combination of classifiers may vary each time
Cascaded Classification	Serial arrangement of trained classifiers	Suitable when the distribution of positive and negative examples is highly skewed
Ensemble Classification	Parallel arrangement of trained classifiers	Inappropriate for scarce datasets
Out-of-distribution approaches	Discarding of input examples of different distribution from the training data	Leads to data loss due to omission of samples. Inappropriate for scarce datasets

Table 3.1: Approaches to robust classification

data. The generation of statistically valid evaluation metrics is also vital and for this, the use of statistical hypothesis testing is key as it is agnostic of the classifier’s estimation and learning process. Based on the above, a methodology that respects the above is proposed in the next chapter, Chapter 4. Classifiers learned with this methodology, can still be deployed in hierarchical, cascaded or ensemble layouts. The proposed methodology does not claim to outperform previously discussed learning methodologies, it proposes a learning framework for heterogeneous and scarce data that do not lay in the unconstrained euclidean space.

Elisavet Charalambous

Chapter 4

A Statistically Unbiased Classification Methodology for Robust Classification

Chapter 2 defined classification and information related to common assumptions and overlooks. Chapter 3 then disclosed related work and practices followed in research with emphasis on the need for standardised methodologies that allow the learning of robust classifiers. Such methodologies are of particular importance due to the diverse applicability of ML methods in real environments and the immense interest of the research community to invent new classification approaches.

The reliability and plausibility of the classification result is dependent on whether imposed assumptions have been respected during learning and the characteristics of the training dataset. Datasets are often scarce, suffering from class under-representation and in a constrained feature space where practices designed for the Euclidean space are not suitable. Nowadays, space constrained data are at least frequent. Common examples represent digitized measurement of analogue representations that exist in the real world. This is because the instrument used for the digitisation process supports a finite range or fidelity. As a result, data emerging from microphones, cameras or even instruments that measure the chemical concentration or composition of artifacts do not lay in the unconstrained real space. It is also important to note that data transformations may be used to transform data from one space to another, however this is not advised for scarce data with uncertainty as their internal structure may be altered. Rather, the use of appropriate distance metrics is advised.

Since the occurrence of space constrained data is common, it was of our interest to develop methods to allow the learning of classifiers that produce plausible and

statistically valid results. Having identified the rather challenging process of reliably categorising data under uncertainty, in this chapter a systematic methodology aiding to solve two major problems in the area is being proposed. First, the robustness evaluation of classification algorithms in successfully categorising a set of data by examining the degree of similarity between discriminated types measured and secondly the ability to evaluate if an expert's labeling (which takes into account a number of attributes) is validated solely by the underlying structure of the obtained data with the deployment of an ensemble strategy.

The above mentioned objectives are based on the development of a statistically valid unbiased methodology that allows the learning of classifiers with multiple algorithms. The suggested methodology is algorithm agnostic in the sense that algorithms are learned and evaluated with the same training and test sets in each iteration. Flexibility also exists in the fine tuning/ parametrisation of each employed classification algorithm with appropriate strategy. The claim of unbiased learning lies on the fact that in each learning iteration, all algorithms are provided with the same data for fine-tuning, training and evaluation. This data driven approach provides freedom in the estimation of algorithm parameters depending on their characteristics and complexity. In each iteration, the validity of the produced classification, on the testing dataset, is examined and the degree of similarity between discriminated types measured. Once a model has been validated for its robustness the domain expert should be able to input new samples and obtain as output the nearest class or classes based on its internal characteristics.

Having identified the challenges in the field and evaluated already deployed practices, a methodology has been implemented for the statistically unbiased classification of scarce heterogeneous data with inherent uncertainty. The proposed methodology has been validated through two independent case studies analysed later in this thesis in Chapters 5 and 6 for the analysis of ceramic archaeological data in the form of chemical compositions and the analysis of audio for the detection of acoustic events in the field of security. Audio samples and the chemical composition of archaeological data represent space constrained examples of data where their transformation from the analogue to the digital realm introduces space restrictions. Additionally, the analogue origin of data also introduces the characteristic of overlapping between different classes. This is because in nature abrupt changes rarely occur and rather transitions between one state to the other are observed.

4.1 Dataset Characteristics

The proposed methodology is particularly suggested to scarce heterogeneous datasets with uncertainties, however, its application is still possible in other types of data. Datasets that find application with this methodology need to be annotated, with hard labels and a one-to-one strict relationship for each input-output pair. This is to say that all samples need to map to one t value within the nominal set of possible classes.

Data scarcity implies that too few data points for each class/cluster are available. This is often because it is difficult to get data or the data is small compared to the amount needed. It is often the case that not all data are sampled for analysis. This is especially observed when other types of analysis are followed to drive the inference process (i.e. utilise knowledge that resides outside the sampled dataset itself). A good example for scarce data is archaeological data where the obtained dataset only forms a small part of the data the archaeologist gathered.

Another characteristic of the data of our interest is heterogeneity. When heterogeneity is combined with scarcity then the effect of sparsity is also observed. Data sparsity refers to cases where data are distributed sparsely over the available space. Data sparsity is the normal situation in all analytics problems as usually only a very small portion of the available state space is filled.

Data scarcity and sparsity often result in highly overlapping classes that do not allow for separable clusters. Depending on the application domain, the overlapping between classes may indicate non-discrete transitions caused by parameters not adequately captured in the dataset (i.e. technological or chronological transitions). Moreover, an aspect that impacts the selection of the appropriate data analysis algorithm is the algorithm's ability to use an appropriate distance metric when similarity or dissimilarity is measured between data samples.

The complexity and dimensionality of the data of interest make the use of clustering methods particularly difficult as the characterisation of samples will be based solely on their internal structure is practically impossible (especially considering that the expert may not annotate data with confidence). Clusters can differ in terms of their shape, size and density. The presence of noise in the data makes the detection of the clusters even more difficult. As a result, the development of classifiers capable of approaching the decision making process of the expert is of interest.

4.2 Preliminary Design Process

The designation of a process capable of fulfilling the above-mentioned objectives required thorough analysis of each influencing parameter. This process was facilitated in iterations where each identified characteristic in the data was analysed separately to allow for enough understanding. One of the primary questions was to determine whether the internal structure of the data is enough to validate the annotations often provided by experts. This process involved analysis of datasets with internal and external validity indices in a less constrained clustering problem.

The process of discriminating unlabeled data seeks solution to two problems. The first, is concerned primarily on the clustering approach and involves: assessing cluster tendency, partitioning and cluster validity. In other words, one should first determine the number of clusters present, then determine which objects belong to each one, and to what degree, and finally validate the how good is the partitioning. Assessing cluster validity is of great importance and the performance of clustering methods greatly depends on specifying the parameters correctly. While the second problem is concerned purely with the way similarity between the different compositions in X is measured. An ideal cluster can be defined as a set of points that is compact and isolated [130].

Possible solutions to the clustering problem requires an integer number c representing the number of clusters which can be either crisp or fuzzy partitions. Crisp clustering can be formulated, in general, as a problem of partitioning the finite set X into a given number c of disjoint clusters and was previously introduced in Section 2.1.2. This definition is also extended to the concept of fuzzy clustering where the principle of partial membership is introduced. That is, due to uncertainties about the integrity of an artifact, errors caused due to the deterioration of materials or other analytical reasons, an artifact may simultaneously belong to more than one clusters. The constraint that each object has unit membership through the total number of clusters needs to hold (i.e. the sum of each row in the U membership matrix should be equal to 1). Given the set X of samples, assign each artifact x to one or more clusters while also specifying the degree of membership for each assignment; this represents the likelihood of the artifact x_i to belong to that specific cluster (see equation 4.1).

$$M_{fc} = \{U \in \mathfrak{R}^{c \times n} | u_{ij} \in [0, 1] \forall j, i; \quad 0 < \sum_{i=1}^n u_{ij} < n, \forall j; \sum_{j=1}^c u_{ij} = 1, \forall i\} \quad (4.1)$$

4.2.1 Clustering Algorithms

For this analysis, k-means and two of its variant algorithms were employed: K-means, Fuzzy c-means and Kernel Fuzzy c-means. Additionally visual clustering is used, as an alternative method, to visually study the characteristics of clusters. Even though k-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering mainly due to ease of implementation, simplicity and efficiency. However, the algorithm is sensitive to the selection of the initial partition and may converge to a local minimum; finitely many possible clustering results exist. Numerous variants of the original approach have been developed with the aim to improve specific aspects of the algorithm and subsequently its effectiveness on specific problems. Visual clustering, on the other hand, is an innovative and alternative approach of clustering multivariate data. Visual Assessing Tendency (VAT) is one of these methods and relies on the very simple principle that similar objects should be placed near each other. The output of the algorithm allows the visualization of similarities between artifacts through a greyscale image.

K-means

Given the set of objects x with n dimensions, the goal is to partition the data in the n -dimensional space into c clusters such that the objective function J has an optimal (usually minimal) value; in hope that the final clustering reflects the structure of the data.

$$J = \sum_{i=1}^c \sum_{j=1}^n u_{ij} d^2(x_j, v_j) \rightarrow \min \quad (4.2)$$

where u_{ij} signifies the membership of object o_i in cluster j , d is the distance metric (defined as in Eq. 4.3) and v_j is the center of cluster j . Inputs to the algorithm are the set of n -dimensional vectors $\{x_1, x_2, \dots, x_n\}$ as well as the parameter c which signifies the desired number of clusters. The algorithm's output is a mapping of the vectors into c clusters (disjoint subsets).

$$d^2 = \|x_j - v_j\|^2 = (x_j - v_j)^T (x_j - v_j) \quad (4.3)$$

The idea behind the operation of this approach is that elements should belong to their closest cluster. The clustering operation terminates when the changes from iter-

ation to iteration fall below the pre-specified positive threshold; usually the threshold δ is equal to $\delta = \frac{0.001}{0.01}$. The algorithm is as follows:

Fuzzy c-means

Fuzzy c-means is an evolution of k-means and incorporates the principle of partial membership allowing data samples to belong partially to more than one clusters. Set X is grouped into c clusters with every data-point in the dataset belonging to every cluster by a certain degree. A data-point lying close to the centre of a cluster will have a high degree of belonging to that cluster and a lower membership when lying further away. This principle allows the algorithm to model in some means the uncertainty in unsupervised learning.

$$J_m = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|x_k - v_i\|^2 \quad (4.4)$$

where m is the fuzzifier determining the level of cluster fuzziness and v the set of cluster centers or prototypes ($v_i \in \mathfrak{R}^p$).

$$\sum_{i=1}^c u_{ik} = 1 \quad (4.5)$$

where u_{ij} signifies the membership of object x_i in cluster j , $\|\cdot\|$ is the euclidian norm and v_j is the center of cluster j .

The cost function of the algorithm is almost identical to the one already presented in k-means since the objective is to minimize the within clusters Euclidean distance. The only difference is that since we do not have crisp clustering the membership matrix U takes floating point values where each column sums up to 1 (the total probability of a sample is 1). Inputs to the algorithm are the set of n -dimensional vectors $\{x_1, x_2, \dots, x_n\}$ as well as the parameter c which signifies the desired number of clusters. The algorithm's output is a mapping of the vectors into c clusters (disjoint subsets).

Kernel Fuzzy c-means

Another variation of k-means is the Kernel FCM variation which additionally to the implementation of the fuzzy partition principle the calculation of the distance between objects is used based on the kernel method; the euclidean distance does not find use in this method. Considering the operation of FCM, KFCM defines a nonlinear map as $\Phi : x \rightarrow \Phi(x) \in F$, where $x \in X$ where X denotes the data space, and F the transformed

feature space with higher even infinite dimension. KFCM minimizes the following objective function:

$$J_m = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|\Phi(x_k) - \Phi(v_i)\|^2 \quad (4.6)$$

Where:

$$\|\Phi(x_k) - \Phi(v_i)\|^2 = K(x_k, x_k) + K(v_i, v_i) - 2K(x_k, v_i) \quad (4.7)$$

in which $K(x, y) = \Phi(x)^T \Phi(y)$ is an inner product kernel function. [61]. In fact, equ. 4.7 can be viewed as kernel-induced new metric in the data space (equ. 4.8), which is defined follows which in turn yields equ. 4.9 :

$$d(x, y) = \|\Phi(x) - \Phi(y)\| = \sqrt{2(1 - K(x, y))} \quad (4.8)$$

The original FCM uses the probabilistic constraint that the memberships of a data point across classes sum to one. While this is useful in creating partitions, the memberships resulting from FCM and its derivatives, however, do not always correspond to the intuitive concept of degree of belonging or compatibility

$$J_m = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|\Phi(x_k) - \Phi(v_i)\|^2 + \sum_{i=1}^c \eta_i \sum_{k=1}^n (1 - u_{ik})^m \quad (4.9)$$

Kernel functions must be continuous, symmetric, and most preferably should have a positive (semi-) definite Gram matrix [220]; this ensures that the optimization problem will be convex and solution will be unique. Choosing the most appropriate kernel highly depends on the problem at hand - and fine tuning its parameters can easily become a tedious and cumbersome task. A polynomial kernel, for example, allows us to model feature conjunctions up to the order of the polynomial while a radial basis function allows to pick out circles (or hyperspheres) [235].

Visual Clustering

The visual assessment of tendency (VAT) technique [27] uses a visual approach to find the number of clusters in data. For object data, visual clustering was initially performed by inspecting scatter plots in $p = 1, 2,$ and 3 dimensions. For $p > 3$, scatter plots cannot be made; not possible for all types of data.



Figure 4.1: The unordered and ordered D matrix: the input and output of VAT

The VAT algorithm reorders the rows and columns of any $n \times n$ scaled dissimilarity matrix D with a modified version of Prim's minimal spanning tree algorithm [198]. Any reordering of D is denoted as D^* . In Figure 4.1 on the left is the un-ordered distance matrix and on the right the ordered version of the same matrix. If the image $I(D)$ has c dark blocks along its main diagonal, this suggests that D contains c clusters. The size of each block may even indicate the approximate size of the suggested cluster [27]. Hence, VAT images suggest both the number of and approximate members of object clusters.

The output of VAT (the ordered D^* matrix) serve as the input to the CLODD algorithm which will determine the number clusters in the data. Specifically, the goal is to partition the objects underlying D and D^* by optimizing an objective function designed to extract aligned clusters from the dark blocks in the image of the ordered dissimilarity matrix $I(D^*)$ [115]. CLODD is a completely autonomous method for determining cluster tendency, extracting clusters from the ordered dissimilarity data, and providing a cluster validity metric. This leads to a distinct advantage of CLODD since it is not tied directly to any distance metric or reordering scheme [115]. The only input requires by the algorithm is an image of reordered dissimilarity data, such that the clusters appear as dark blocks along the diagonal.

4.2.2 Experimental Results

It was of interest to evaluate the operation and effectiveness, of the introduced algorithms, in one of our domains of interest. For this purpose, an archaeological dataset

of chemical compositional data was used.

The Dataset

The study involved the compositional analysis of a small dataset obtained from the ED-XRF (Energy Dispersive X-Ray Fluorescence) analysis of ceramics [70]. A quite challenging case since the dataset is consisted of 73 samples - from which 6 were clearly noted as outliers - classified by the expert archaeologist into 10 classes, the objective of analysis was the diachronic technological assessment of pottery production and patterns of ceramic distribution at a single settlement; in an attempt to reconstruct production and distribution patterns. The selected samples (are thought to) represent the main wares and document technological differences in each of these wares over time.

The dataset became subject to treating before statistical analysis was performed. This included the conversion of all elements into oxide compounds with stoichiometry. The composition of each artifact was then normalized (i.e. force the sum of each row to be 100). Also, It is typical with archaeological data to exclude certain features upon processing. Trace elements with elemental concentration below 10 ppm were omitted along with sulphur trioxide (SO₃), cerium oxide (CeO₂), Chlorine (ClO) and lead oxide (PbO) concentrations due to analytical reasons imposed by the instrument's inability to provide measurements. Sodium oxide (Na₂O), phosphorus pentoxide (P₂O₅), sulphur trioxide (SO₃), cobalt (Co₃O₄) and cerium oxides (CeO₂) were also omitted from multivariate statistics because of their inconsistent values and poor reproducibility in successive analytical runs [70].

The obtained dataset forms only a small part of the data the archaeologist gathered. ED-XRF analysis was performed as a complementary analysis method so as to confirm and refine the interpretation of inferences that were obtained after the application of traditional approaches. So as to allow the evaluation of the correspondence between the mineralogical and chemical groupings and define their degree of consistency. An extensive description of the original study along with comprehensive explanations of the processes that were followed by the archaeologist can be found here [70].

The Approach

The aim of the experiment was to obtain quantitative measures on each method's performance so that comparable results could be obtained. The small size of the

dataset made necessary the use of a re-sampling method so as to allow the generation of statistical figures. Bootstrapping with replacement was used and multiple new datasets were generated based on the original one.

Since, the artifacts were measured for both main and trace elements, the resultant data is consisted of features with different scales. Due to this, feature/column-wise pre-processing was necessary. Each data point in the matrix is divided by its column arithmetic mean. This quite simple transformation allows the relative variability in the original variables to become the variability of the transformed ones [254]. As a consequence of this transformation, the original variables become comparable with each other. The arithmetic mean of each new variable is equal to 1. The variables preserve their different original variability and therefore have different influence on the cluster analysis [74].

Once clustering was performed by each, of the already discussed, clustering algorithms, a series of metrics were calculated with the aim to estimate the number of clusters. This involved an iterative method in which each algorithm was called to performed clustering for different values of c . Knowing that the data were classified by the expert into 10 (however, non-definite) classes, the range of c selected to be $c=1, \dots, 12$.

Upon each iteration, and since we assume that the true label of each object is known, the external validity indices: Rand Index [203], the Adjusted Rand Index [127], the Mirkin Index [172] and the Hubert Index were calculated for each clustering [167].

In short, the Rand Index penalizes both false positive and false negative decisions during clustering. The adjusted Rand index is the corrected-for-chance version of the Rand index. Though the Rand index may only yield a value between 0 and +1, the adjusted Rand index can yield negative values if the index is less than the expected index. The Mirkin Index, corresponds to the hamming distance between certain binary vector representations of each partition; it yields 0 for identical clusterings and positive otherwise [172]. Finally, the Hubert's Index has a clear probabilistic interpretation and is corrected for chance with respect to the null hypothesis, and is bounded between -1 and +1.

The figures which follow (Fig. 4.2- 4.5) show the results of calculating the above indices for each clustering with respect to the different values of c .

During the processing of each resampled dataset, a matching factor was also calculated. This had to do with the successful assignment of elements to the correct target class. Each set of clusters, produced by the algorithms, was compared to the known

target classes and a percentage of matching elements of each cluster was calculated. For the fuzzy solutions, a crisp membership matrix was required. This involved the assignment of each artifact to the cluster with maximum membership. The calculation of the matching metric was performed only at the solutions that yield a number of clusters equal to the number of the known target classes (i.e.10). An overall matching percentage for each algorithm was calculated, by averaging the values of all iterations.

K-means and Variants

The evaluation of the k-means method was straightforward since it does not require other parametrization than defining the number of clusters. However, this is not the case for the FCM and KFCM algorithms which require the analyst to determine the values of other operational parameters. Both methods perform fuzzy clustering and therefore require the specification of the fuzzifier parameter. Even though the fine tuning of the parameter has a significant impact on the performance of the algorithm. This value was chosen to be $m = 2$; a reasonable selection based on literature. This fixed value was adopted for both algorithms so as to avoid the optimization of their performance to this specific dataset.

KFCM also requires the analyst to deliberately select the type and parameters of the kernel to be used [126]. The algorithm was evaluated for two different types of kernels, a second order polynomial ($K = (XX' + 1)^2$) and the RBF kernel $K = e^{-\frac{1}{(2\sigma^2 XX')^2}}$ with a σ value of 1.5).

The figures (Fig. 4.2- 4.5) illustrate that the fuzzy solutions produce more stable results than the crisp k-means (Fig, 4.2). This result was expected; the archaeological data is consisted of non separable clusters not only due to the similarities between the different artifacts but also due to the fact that artifacts span chronologically many centuries. The performance problems of the algorithms become apparent with the calculation of the adjusted Rand index; consistently below 0.6 with FCM being the method which produces the highest values. The performance of the algorithms against the adjusted Rand index is of great importance due to its sensitivity and ability to not be affected by the granularity of each particular clustering.

The matching percentage of the k-means algorithm was consistently around 33.5-35.2% while FCM tended to produce higher percentages of the order of 47.5-48.75%. The case of the KFCM algorithm was quite interesting. The performance of this ap-

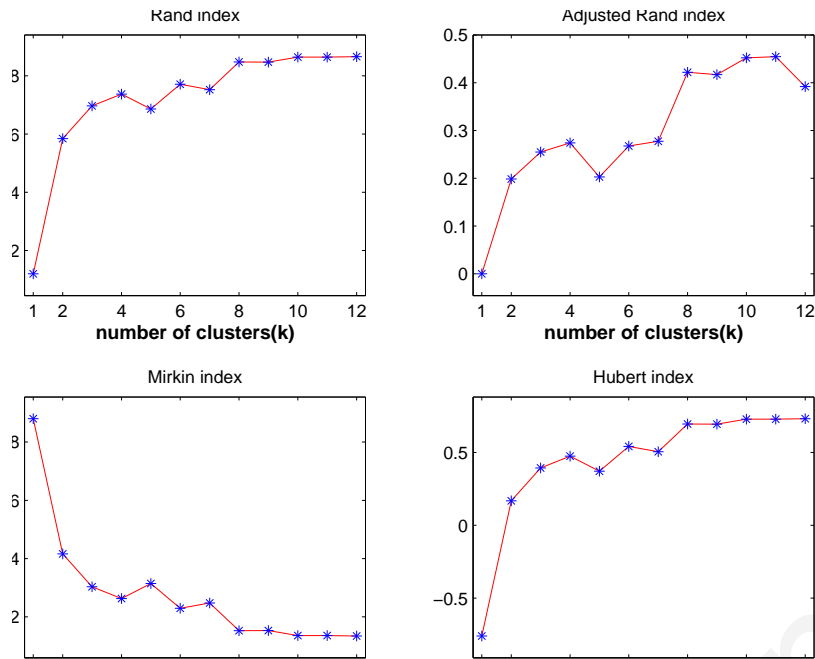


Figure 4.2: K-means: External Validity Indices

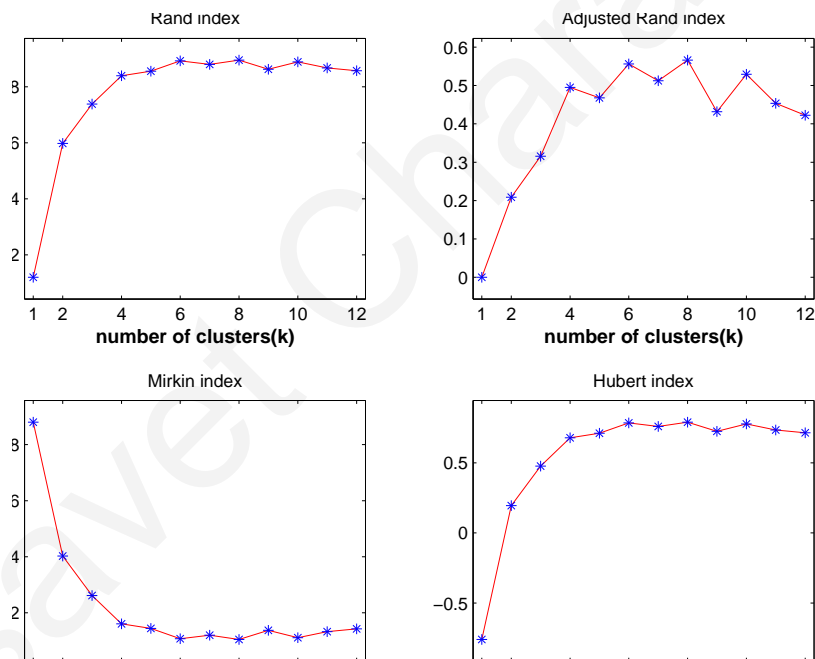


Figure 4.3: FCM: External Validity Indices

proach was significantly affected by the kernel that was used and consequently its parameters. KFCM with the polynomial kernel produced a matching percentage between 43-48% while the RBF kernel returned 19.2%, a significantly lower percentage. Overall, less than 50% of the artifacts were assigned to the correct clusters.

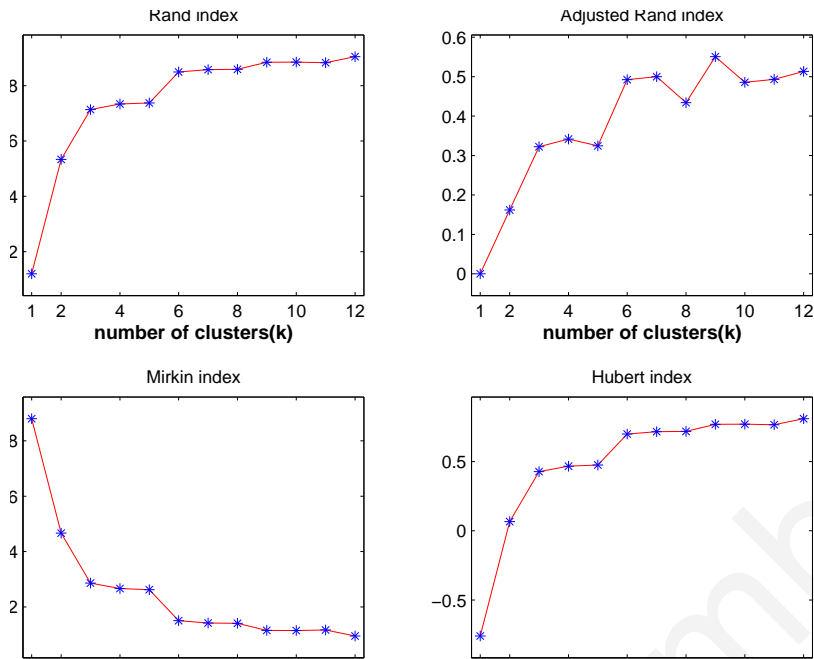


Figure 4.4: KFCM Polynomial Kernel: External Validity Indices

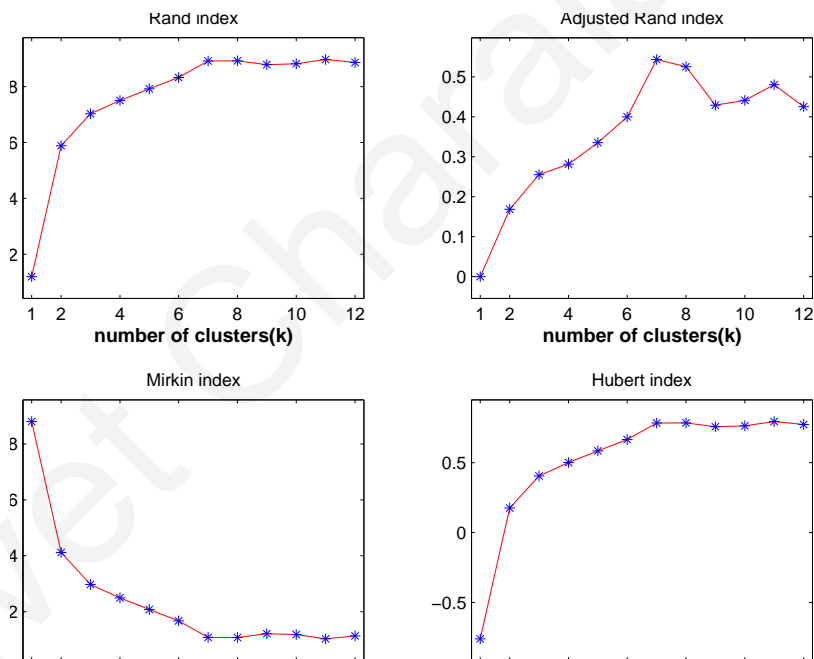


Figure 4.5: KFCM RBF Kernel: External Validity Indices

VAT Clustering

Since the VAT and CLODD algorithms are non-parametric and only accept as input the symmetric distance matrix of the objects in set X . For the needs of the experiment the distance matrix of the data served as the input to the VAT algorithm which subsequently produced the ordered matrix solution. The output of the VAT algorithm was then used as the input to the CLODD algorithm which produced the estimate of the

number of clusters.

The evaluation of the algorithm against the external validity indices was performed for all datasets including the original one (Fig. 4.6). The VAT algorithm, according to the values of adjusted Rand index, managed the best results, out of the algorithms we have evaluated, by reaching values sometimes higher than 0.7. Additionally, the CLODD algorithm produced a mean number of clusters: 9, a quite impressive result approaching the true value especially considering the complexity of the data.

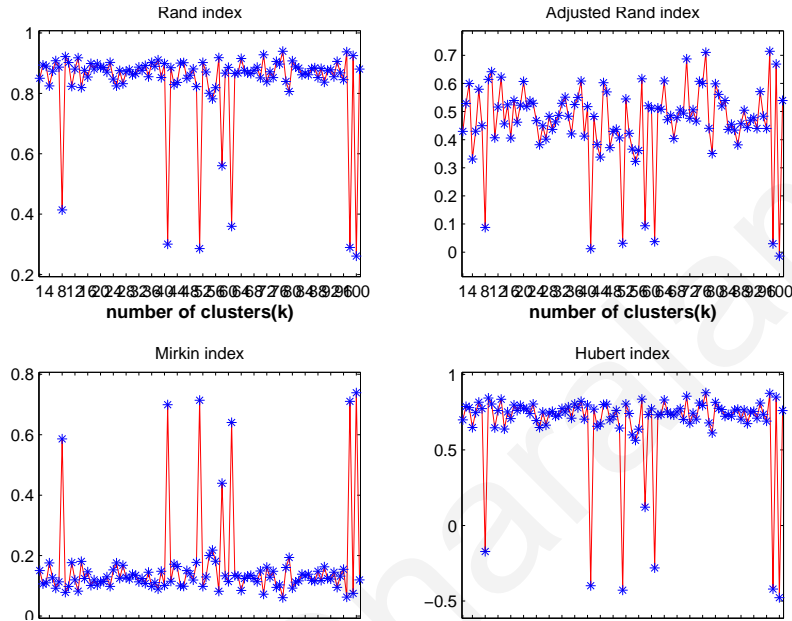


Figure 4.6: The effectiveness of CLODD for different datasets

4.2.3 Key Findings

Scarce and heterogeneous data can be difficult to analyse as they often suffer from class/cluster overlapping. This task becomes even harder when data are subject to uncertainty. Example data that adhere to these characteristics involve chemical compositional archaeological data from pottery.

Analysis of the sample dataset revealed that fuzzy clustering solutions seemed capable of capturing amounts of data uncertainties. However, it should be considered that evaluation based on external validity metrics required hard partitioning of the data which resulted in altering the original fuzzy solutions which is an undesirable effect.

Evaluation was only allowed with the generation of new datasets based on re-sampling method with replacement. Considering the size of the dataset this may result in the exclusion of some classes or the selection of very few samples out of certain classes

and therefore the inability of the clustering techniques to identify the different groups.

The application of visual clustering techniques on archaeological data produced interesting results as it indicated a number of possible clusters close to what the expert suggested. Additionally, visual clustering also provided useful information with respect to the size of clusters and their in-between distance. Visual clustering indicated relationship between clusters. However since the output of VAT is unordered it was of interest to further experimentation with the generation of confusion matrices.

The key findings of the discussed study served as the foundation for the development of a methodology for the learning of robust classifiers in a algorithm agnostic approach. In the rest of this chapter we will be discussing the implementation details of this approach and the reasoning behind the selection of the employed methods.

4.3 A Methodology for Robust Classification

The study in the previous section highlighted the use of re-sampling along with the use of the appropriate distance metrics impact the classification outcome. As a result, this section suggests a design that follows a systematic approach and well-established methods, such as bootstrapping with replacement for dealing with data scarcity and uncertainties and the 5×2 cross validation (paired t-test and F-test) tests, to ensure that the results are statistically significant, whereas classification evaluation is performed on the basis of an appropriate classification metric. The classification evaluation is measured with the multi-class ROC AUC evaluation metric as it represents a degree or measure of separability between classes.

The developed strategy may be applied to both multi-class and binary classification problems with labeled datasets. Data annotation assumed to be performed by experts in each respective field with the utilisation of knowledge not solely emerging from the dataset itself. Due to the scarcity of data and the aspect of under-represented classes, outliers are included in the analysis and are considered as single-element-classes.

The aim of the presented methodology is neither to achieve perfect classification, nor to determine the best classification algorithm. The target is rather to develop a plausible, unbiased and statistically valid methodology for classification, which takes into consideration the idiosyncrasies of the classification algorithm, in general, and characteristics of the data and application domain in particular. It therefore allows data analysts to select the best performing algorithm by applying the proposed methodology

with classification algorithms of interest. Pair-wise analysis of selected model's evaluation may be used as part of statistical testing towards the validation or rejection of the hypothesis that the under evaluation trained models perform equally well. Statistical testing allow the analyst to identify that a model's performance is consistently – with small error probability – better than its opponent.

Additionally, the proposed methodology aims to examine the validity of the produced categorisation. The proposed methodology may subsequently be used to differentiate a series of samples, and investigate the degree of similarity between discriminated types. A high level diagram of the proposed methodology is summarised in Figure 4.7 in a number of steps that involve, data boosting with replacement, feature selection, algorithm parametrisation, classification evaluation and hypothesis testing.

4.3.1 The Dataset

The objective of this framework is to learn a mapping from inputs x to outputs t , given a labeled set of input-output pairs $D = (x_i, t_i)_{i=1}^N$ with D being the training set and N the number of training examples. Given a training set of N example input-output pairs

$$(x_1, t_1), (x_2, t_2), \dots (x_N, t_N)$$

where each t_i was generated by an unknown function $t = f(x)$, discover a function h that approximates the true function f so it generalises well when provided with data of same distribution characteristics as inputs x . The outputs t need to be hard labels (no partial labeling or weak labeling).

This methodology is claimed to be algorithm agnostic as it allows the learning of any algorithm, as long as its applicability is possible with the above characteristics.

4.3.2 Significance Testing with Simulation

Statistical testing requires the calculation of enough statistics to alleviate the factor of likelihood in the classification output. Since data availability is an issue in scarce data the drawing of statistics will be performed through bootstrap simulation. During each simulation iteration, 5×2 cv is applied and the classification performance of each algorithm is used to calculate the statistics for significance testing with the t-test and F-test.

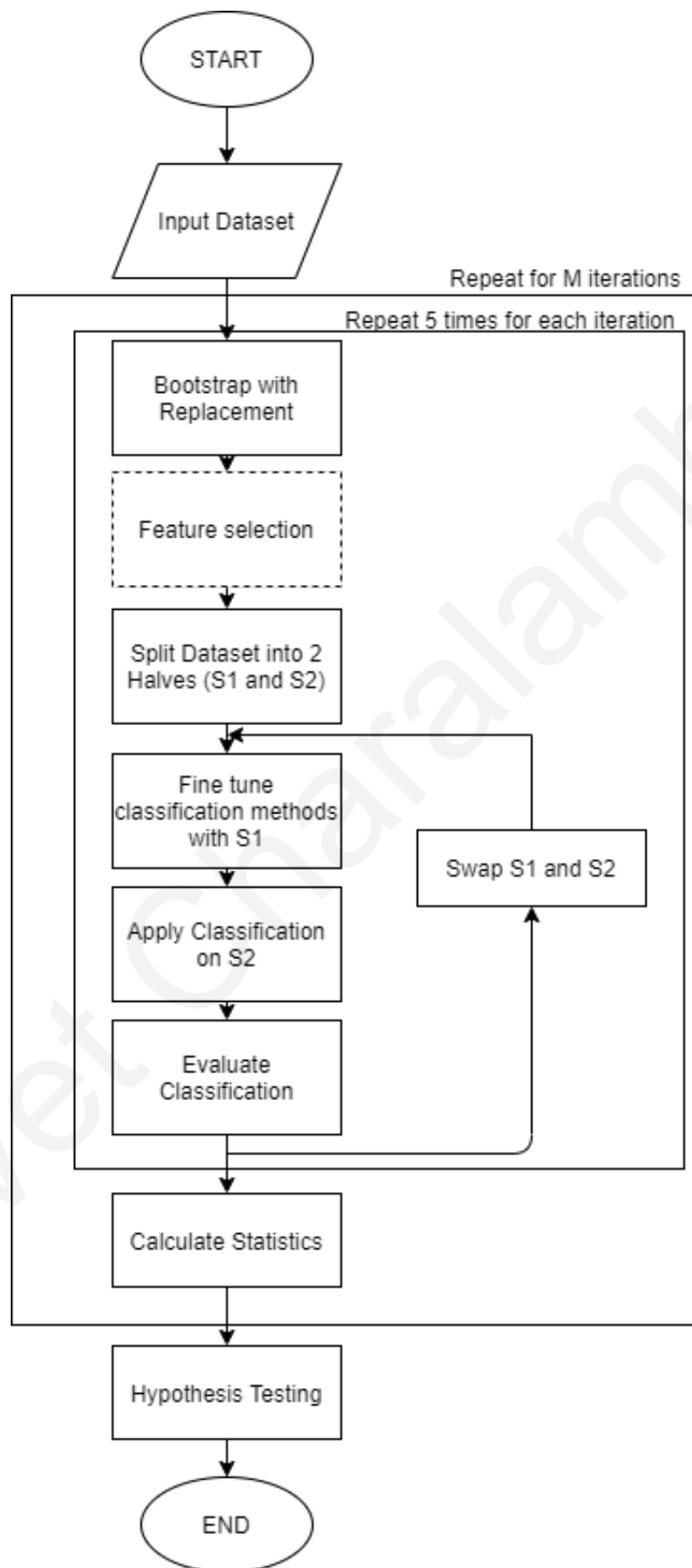


Figure 4.7: Flowchart of methodology

Hypothesis tests on paired data can be analysed by considering the differences between the paired items. The p-statistic produced by each of the under training algorithms are retained to measure the power and significance statistics.

The distribution of differences is usually symmetric, in fact, the distribution must be symmetric if the individual distributions of the two items are identical. In reality, this is not always the case since equal class representation is not ensured when random sampling is applied. Due to this, significance testing methods derived from the paired t-test are selected the paired t-test finds applicability even when the distributions of the individual items are not normal.

The simulation process is iterative and a large enough number (M) of iterations need to be performed. M is a large number, usually anywhere from 100 to 10,000. As the number of iterations is increased, the accuracy and running time of the simulation will be increased also.

The precision of the simulated power estimates are calculated by the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives you an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the 'precision' value given in the table Table 4.1).

Estimate Iterations		
Simulation Size M	Precision when Power=0.50	Precision when Power=0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
5000	0.004	0.002
10000	0.003	0.001

Table 4.1: Estimate Iterations

The power of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is, the power of a hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is the hypothesis that

is true. The power value in a simulation can be estimated as the proportion of times that the null hypothesis is rejected (out of M).

4.3.3 Algorithm Parametrisation

There is no equivalence between the estimation of each algorithm's parameters for classification, therefore, the proposed methodology allows the application of a suitable parameter optimisation strategy for each algorithm. It is of primary interest to learn classifiers that produce minimal mis-classification rate. As a result, the methodology allows freedom in the deployed algorithm parametrisation process.

During each training iteration, the same (re-sampled) training and testing datasets are provided as input to all algorithms. Despite the fact that the training data need to be similar for all algorithms, the parametrisation process might be adapted to the needs and characteristics of the algorithm. This is so algorithms with many parameters are allowed to more extensive parametrisation than algorithms with fewer parameters.

The learning of more than two algorithms is possible, even though paired tests are used for significance testing. During learning, the classification evaluation produced by each algorithm in each $5 \times 2cv$ simulation iteration can be retained and used to estimate the outcome of pair-wise significance tests to validate or reject whether a trained model is significantly better than its opponent.

4.3.4 Classification Evaluation

AUC - ROC, previously introduced in Chapter 3 curve is a performance measurement for classification problem at various thresholds settings. ROC AUC is a metric on how much model is capable of distinguishing between classes, the higher the AUC, the better is the model at predicting samples in their correct class.

In a Receiver Operating Characteristic (ROC) curve the TPR (True Positive Rate aka sensitivity) is plotted in FPR (False Positive Rate), calculated as $1 - \text{specificity}$, for different cut-off points, where TPR is on y-axis and FPR is on the x-axis (shown in Figure 4.8). Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

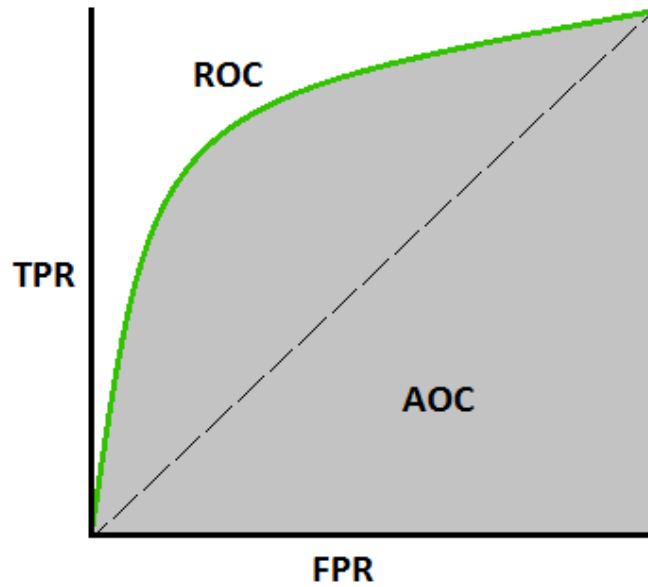


Figure 4.8: AUC - ROC Curve [184]

The ROC AUC metric is the preferred classification evaluation metric due to its direct relationship with the sensitivity and specificity metrics. In Chapter 3 the importance of the sensitivity metric as a form of measuring robustness has been highlighted. Sensitivity measures the proportion of actual positives that are correctly identified as such while specificity measures the proportion of actual negatives that are correctly identified as such.

The use of AUC as a metric for classification evaluation and robustness is important, as it is an algorithm and parametrization agnostic metric that complies with the objectives and aims of the proposed methodology. The selection of this metric emerged as a result of a number of experiments with various classification evaluation metrics, of which some are disclosed in Chapters 5 & 6 through the validation of the methodology in the two case studies.

When using normalized units, the AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming “positive” ranks higher than “negative”) [89]. This can be seen as follows: the area under the curve is given by (the integral boundaries are reversed as

large T has a lower value on the x-axis)

$$TPR(T) : T \rightarrow y(x) \quad (4.10)$$

$$FPR(T) : T \rightarrow x \quad (4.11)$$

$$AUC = \int_{x=0}^1 TPR(FPR^{-1}(x))dx \quad (4.12)$$

ROC curves are typically used in binary classification but it can be extended to multi-class classification problems by binarising the classification output and drawing a ROC curve per label.

4.3.5 Bootstrapped 5×2 cv t-test and F-test

The algorithm learning method should be repeated for enough iterations based on simulation principles, introduced in Section 4.3.2, to remove the bias of variance between iterations. In each iteration, bootstrap with replacement is applied to produce datasets of similar size to the input dataset. This step in the methodology has been added to deal with the scarcity in data. It is expected that classes will not be equally represented in all iterations due to the replacement strategy, however, this effect is alleviated by running enough iterations.

Bootstrap with replacement involves a computer-intensive re-sampling method, introduced in statistics by B. Efron in 1979 [80] for estimating the variability of statistical quantities and for setting confidence regions. This method allows the re-sampling of data – given observations D_1, \dots, D_n , artificial bootstrap samples are drawn with replacement from D_1, \dots, D_n , putting an equal probability mass of $\frac{1}{n}$ on D_i for each $i \in \{1, \dots, n\}$.

Bootstrap is commonly implemented with the paired t-test, however in our design bootstrap with replacement is paired with the 5×2cv t-test and F-test to benefit from their advantages when dealing with small datasets.

Our choice of five replications of cross-validation is not arbitrary. Exploratory studies showed that using fewer or more than five replications increased the risk of type I error. A possible explanation is that there are two competing problems. With fewer replications, the noise in the measurement of the s_i 's becomes troublesome. With more replications, the lack of independence among the s_i 's becomes troublesome. Whether five is the best value for the number of replications is an open question [68].

In each simulation iteration, a new dataset is generated with the bootstrap with

replacement method. The generated dataset is used as input to the 5×2 cv configuration. In the 5×2 cv t-test, proposed by Dietterich [68], we perform 5 replications of 2-fold cross validation. In each replication, the input dataset is re-sampled without replacement, and divided into two equal-sized sets. These sets are usually referred to as training and testing sets in the 2-fold cross validation process. These training and testing datasets will be alternated during the 2^{nd} fold of training (i.e. data samples in the training set will be used for testing and the opposite).

The learning process in each replication in 5×2 cv starts with the parametrisation of each algorithm. For this the, the training dataset from the previous step is again sub-divided in two halves (S_1 and S_2 from Figure 4.7). S_1 is used for the estimation of each algorithms' parameters. The estimated parameters are then used to validate each models performance on S_2 . The estimated values of S_2 are compared to the actual class labels to produce a metric of evaluation of the classification. Datasets S_1 and S_2 are then swapped and the process of algorithm parametrisation is now repeated on S_2 and validated against S_1 .

The described procedure involves one iteration of five in the 5×2 cross validation method. Once all five iterations are obtained, the t and F statistics are measured on the ten validation values to obtain one figure for each of the M simulation iterations. In our proposed framework, the classification performance is evaluated with the ROC AUC metric, while it is advised to deploy an appropriate distance metric should the algorithm allows.

4.3.6 Estimation of t and F statistics for significance testing

The estimation of t and F statistics for significance testing require the calculation of the difference ($p_i^{(j)}$) between the error rates of the two classifiers, for each simulation iteration, on fold $j = 1, 2$ of replication $i = 1, \dots, 5$. The average on replication i is $\bar{p}_i = (p_i^{(1)} + p_i^{(2)})/2$ and the estimated variance is $s_i^2 = (p_i^{(1)} - \bar{p}_i) + (p_i^{(2)} - \bar{p}_i)^2$. Under the null hypothesis, $p_i^{(j)}$ is the difference of two identically distributed propositions so can be safely treated as a normal distribution with zero mean and unknown variance σ^2 . Then $p_i^{(j)}/\sigma$ is unit normal. If $p_i^{(1)}$ and $p_i^{(2)}$ are independent normals, s_i^2/σ^2 is chi-square with one degree of freedom. Then

$$M_s = \frac{\sum_{i=1}^5 s_i^2}{\sigma^2}$$

is chi-square with 5 degrees of freedom. if $Z \sim \mathcal{Z}$ and $X \sim \chi_n^2$

$$T_n = \frac{Z}{\sqrt{X/n}}$$

is t-distributed with n degrees of freedom. Therefore

$$t = \frac{p_1^{(1)}}{\sqrt{M_s/5}} = \frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^5 s_i^2/5}}$$

is approximately t-distributed with 5 degrees of freedom [68]. The hypothesis that the two classifiers have the same error rate with 95% confidence is rejected, if $t > 2.571$. We note that the numerator $p_1^{(1)}$ is arbitrary and actually there are ten different values that can be placed in the numerator, leading to ten possible statistics

$$t_i^{(j)} = \frac{p_i^{(j)}}{\sqrt{\sum_{i=1}^5 s_i^2/5}}$$

The combined $5 \times 2cv$ F-test is a new test that combines the results of the ten possible statistics and promises to be more robust. If $p_i^{(j)}/\sigma \sim \mathcal{Z}$, then $(p_i^{(j)})^2/\sigma^2 \sim \chi_1^2$ and

$$N = \frac{\sum_{i=1}^5 \sum_{j=1}^2 (p_i^{(j)})^2}{\sigma^2}$$

is chi square with 10 degrees of freedom. If $X_1 \sim \chi_n^2$ and $X_2 \sim \chi_m^2$ then

$$\frac{X_1/n}{X_2/m} \sim F_{n,m}$$

Therefore, we have

$$f = \frac{N/10}{M/5} = \frac{\sum_{i=1}^5 \sum_{j=1}^2 (p_i^{(j)})^2}{2 \sum_{i=1}^5 s_i^2}$$

is approximately F distributed with 10 and 5 degrees of freedom. For example we reject the hypothesis that the algorithms have the same error rate with 0.95 confidence if the statistic F is greater than 4.74. The combined version combines the ten statistics and is more robust; it is as if the combined version takes a majority vote over the ten possible $5 \times 2cv$ t-test results.

The values of t and F statistic of each simulation iteration can be used to estimate the power of the test and therefore the model's robustness estimate in classifying new inputs x

4.3.7 Feature Selection

Feature selection may also be applied as an optional step that should be applied in a case by case situation. The use of ensemble feature selection with random forests is suggested to allow for variability when used in combination with boosting. The tree-based strategies used by random forests naturally rank features by how well they improve the purity of nodes. Nodes with the greatest decrease in impurity happen at the start of the trees, while nodes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node creates a subset of the most important features which are retained. The threshold of pruning is a parameter that needs to be determined on a case by case scenario and it is dependent on the dimensionality of the feature space and the degree of overlapping between class distributions (data complexity). The dimensionality of the dataset is reduced to only retained features above the specified pruning threshold.

Random forests (RF) construct many individual decision trees at training. Predictions from all trees are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. For each decision tree, nodes importance using Gini Importance, assuming only two child nodes for simplicity (binary tree):

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

where ni_j the importance of node j , w_j the weighted number of samples reaching node j , C_i the impurity value of node j , $left(j)$ the child node from left split on node j and equivalently $right(j)$ the child node on the right split of node j . The importance for each feature on a decision tree is then calculated as:

$$fi_i = \frac{\sum_{j: \text{node } j \text{ split on feature } i} ni_j}{\sum_{k \in \text{all nodes}} ni_k}$$

, where fi_i the importance of feature i and ni_j the importance of node j . These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$norm fi_i = \frac{fi_i}{\sum_{j \in \text{all features}} fi_j}$$

. The final feature importance, at the Random Forest level, is its average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in \text{alltrees}} \text{norm}f fi_{ij}}{T}$$

where $RFfi_i$ the importance of feature i calculated from all trees in the Random Forest model. $\text{norm}f fi_{ij}$, the normalised feature importance for i om tree j and T the total number of trees.

4.4 Identification of inter-class relationships

The application of the previously described methodology on the same set of data may produce interesting findings with respect to class similarity. Based on the fact that outliers were included in the analysis and represented single element classes, patterns in their mis-classification indicate similarities with other classes. Visualisation of results in the form of a confusion matrix may summarise the prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

Due to the fact that re-sampling with replacement is a core step in the proposed classification methodology, observation of confusion matrices may lead to tracing inter-class relationships. This may be done by superimposing the confusion matrices obtained from the re-sampling iterations. Overlaying the confusion matrices alleviates the factor of misclassification by chance.

Identification of such relationships implies that if a sample from a certain class is to be misclassified, it would be classified into a very small number of candidate classes. Inter-class relationships might reveal similar characteristics between samples of specific classes. However, when inter-class relationships are found between classes and single class elements, this might indicate that the specific sample is wrongly misclassified as an outlier.

4.5 Summary

The supervised classification problem may be broken down into two separate stages: 1) the inference stage where training data is used to learn a model and 2) the decision

stage in which the trained model is used to make class assignments [29]. The design of the methodology to tackle a problem has to serve its specific needs while considering limitations and constraints. The choice of deployed methods accounts for the non-uniformity in the distribution of classes, as well as the size of the dataset, which might be relatively small compared to the number of classes.

Classification of small datasets (with regard to the number of classes) requires the use of re-sampling methods. For the needs of this study, bootstrapping with replacement is used to allow the generation of datasets of original size. The choice of re-sampling. adds the assumption that the dataset is representative to the population, in other words, the sample of collected ceramics yields a good representation of the population.

In many cases, probability theory is applied to problems that involve uncertainty where the term uncertainty is related to the prediction of an item's class based on prior knowledge, the fit of the model to the given data, the evaluation process of the deployed model as well as on attributes that account for an expert's uncertainty in categorising an item or a measuring instrument's ability to record/collect data measurements with precision and accuracy.

Even though, not formally introduced as integral part of the inference stage, feature selection is a highly recommended technique in highly dimensional feature spaces. In many problems, it is common for features not to follow a normal distribution also causing skewed distributions dominating other key features. Feature selection is therefore a technique that allows identifying the most "important" features in data; the ones that contribute most to the target variable. In other words, feature selection allows an expert to choose the best predictors for the target variable, these are the features that allow maximum discrimination. Often in data science experts are encountered with data with hundreds or even millions of features and are called to create a model that only includes the most important features. In this problem, feature selection allows the training of models that are simple to interpret, reduce the variance of the model – and therefore lessen over-fitting – and finally reduce the complexity of the model in terms of computational needs as far as resources and time are concerned. Using less redundant data in the analysis also lessens the possibility of making decisions on redundant data/noise possibly leading to improved accuracy.

Despite the benefits, one of the main drawbacks of feature selection is the assumption that the input dataset is representative of the population. Additionally, in highly

complex data (where class overlapping and similarity between clusters exist), removing features might introduce adverse implications. The use of feature selection should be implemented on a case by case manner and may therefore serve as an optional step. In each case the expert needs to investigate whether feature selection improves the classification result or not. An example of highly complex data will be presented in Chapter 5 where main and trace elements contribute almost equally to class discrimination.

The deployment of established methods allows the evaluation of the validity of the results through the use of a special form of cross validation testing. The developed design follows a systematic approach and well-established methods, such as bootstrapping with replacement [80] and the 5×2 cross validation (paired t-test and F-test) tests in order to ensure that the results are statistically significant. The learning process involves aspects that introduce randomness. One may claim that this process may favor one algorithm over the other. However this factor is alleviated by allowing enough iterations during the simulation where a large enough number of iterations should be allowed.

Moreover, the performance of the algorithms is evaluated using classification evaluation metrics. A range of evaluation metrics exist. Despite the fact that the purpose of these metrics is to evaluate how effective a trained model is at classifying correctly the input data, they account for different factors. The selection of the appropriate evaluation metric depends on the nature and cost factors of the problem as well as on the type of the selected classification algorithm. An analysis on the impact of classification evaluation metric has been performed to investigate the degree of impact in the obtained measurements. The results of this study revealed that the use of the appropriate evaluation metric will affect the selection of the more effective classifier; the use of an inappropriate metric will hinder a classifier's performance if the necessary factors are not accounted for (i.e. categorise samples to the exact category, minimise the number of false positives, minimise the number of false positives).

Since the performance of any algorithm is highly dependent on its parameterisation, it is necessary to consider some fine-tuning of each algorithm's parameters for each bootstrapped dataset. This operation however may not be considered equivalent in every algorithm as each algorithm's parameters cannot be assumed to be identical. For instance, in [46] the C4.5, k-NN and LVQ algorithms were deployed. Since the LVQ algorithm requires the parameterisation of more parameters than the other two, it is believed that its performance is more likely to suffer if not exhaustive fine-tuning is

performed. In this study, the performance of the algorithms was evaluated against the classification accuracy and the Jaccard Index. Even though the classification accuracy is the predominantly used metric, it does not account for separability in the data. As a result, further investigation was performed with respect to the most appropriate classification evaluation metric.

The application of the proposed learning framework is further discussed in Chapter 5 for the analysis of scarce chemical compositional data that emerge from archaeological pottery and in Chapter 6 for the analysis of scarce audio data to achieve acoustic event detection as part of a security surveillance system.

Chapter 5

Case study I: Analysis of Compositional Archaeological Data with Uncertainties

5.1 Introduction

The perceived present is the consequence of human action in the past, interacting with natural processes through time [16,205]. Archaeology ultimately aims at investigating social causation through the examination of gathered residue evidence [17]. Pottery analysis, in particular, has been proven cross-culturally an indispensable tool for indirectly approaching past people and societies. For this reason, compositional (mineralogical and chemical) and micro-structural analyses have become an integral part of interdisciplinary archaeological research, underlining the importance of compositional and technological comparative studies [46].

Archaeological questions, for which ceramics are used, vary from straightforward analytical questions (i.e. identification of distinct (chemical) groups within the data and association with different origins or manufacturing technologies) to behavioural analytical questions (i.e. the investigation of the relationship between the "recipe" and the sources of raw materials). As a consequence, classification analysis of archaeological ceramics is being deployed by researchers to provide answers mainly to analytical questions through exploratory analysis of chemical compositional data with the aim to identify clusters. However, compositional data impose a number of restrictions during data analysis. An example involves sub-compositional coherence and the fact that features do not vary independent of each other. This property requires the selection of a distance metric such that the distance between samples does not decrease as the

number of species considered increases, where use of ratios of components.

The key challenges of archaeological data analysis are the small number of samples, the uncertainty in the compositional measurements as well as the uncertainty in the expert classification. These impose challenges in applying Machine Learning methodologies for classification purposes. Classification of archaeological ceramics deals with the categorisation of ceramic specimens of similar chemical profiles [179]. Classification in archaeology is very important since it makes possible the identification of a newly found artefact based on already known information.

The methodology introduced in Chapter 4 is applied in the analysis of chemical compositional data with the aid to solve two major problems in the area. Firstly, the performance evaluation of classification algorithms in successfully categorising a set of data. Secondly, the ability to evaluate if an expert's labeling (which takes into account a number of attributes) is validated solely by the underlying structure of the compositional data.

Archaeological artefacts and ceramics in particular, constitute a class of data notably challenging for analysis exhibiting characteristics such as uncertainty in the data measurements, due to the natural deterioration of materials and inconsistencies in deployed analysis methods/instruments as well as uncertainty in the label due to the expert's low confidence during data annotation caused for a number of reasons (disparity in data, artefact condition and composition, analysis method, incomplete data etc.); data scarcity and heterogeneity in composition also contribute to this.

Another common attribute of archaeological artefacts is the overlapping between classes due to (possibly) mutual sources of raw material. The non-separability of samples might be explained as a relationship between classes adding to inference conclusions by the expert.

Machine learning in the field of archaeology has been used mainly as a method for performing exploratory analysis using only part of the available information (chemical, compositional, mineralogical, macroscopic etc.) while archaeologists currently cannot rely on structures emerging from ML methods due to doubts in their reproducibility and reliability. Therefore, our research is focused on chemical heterogeneous archaeological data where the composition of the sample under analysis is being expressed in chemical element compositions.

5.2 Compositional Archaeological Data under Uncertainty

Archaeological data constitute a special class of data exhibiting characteristics imposing challenges during classification. Archaeological artefacts not only become subject to natural deterioration over the years but also candidate analysis samples for a number of methodologies (and technologies), introducing uncertainty. It has also been noted that at times not even the expert may categorise artifacts with confidence and therefore data analysis methods need to consider this [46]. Additionally, the field of archaeology suffers a lot from the issue of data scarcity where not sufficient data are available for their reliable and robust characterisation [268].

Many parameters influence the reliability of the produced data. Different people execute the same procedures in different ways; thereby increasing the within class variance. This problem gets even worse by taking into account that apart from variations generated due to the human factor, acquired variability is also caused due to the deterioration of the source material because of its natural ageing as well as the environment of preservation.

The composition of a sample might be characterised as homogeneous or heterogeneous. With respect to chemical compositional analysis, homogeneous refers to the situation where a sample has uniform composition and properties throughout while heterogeneous mixtures have particles that can be seen under a microscope. Heterogeneous mixtures are jumbled irregularly together and as a result can usually be separated in two or more homogeneous mixtures [208]. In archaeology, homogeneity and heterogeneity commonly also imply situations where certain features dominate the analysis; the most highly variable elements have the greatest of the impacts on the multivariate data ensemble and that they do not necessarily depict elements with high concentrations [204]. This is particularly visible in the analysis of ceramics where just a few chemical elements may constitute 90-95% of an artefact's composition however it's been believed that in ceramics, the inclusions are of significant importance during categorisation by experts [105].

Data annotation is an additional, expensive, and error-prone preparation process. Individual data have to be carefully inspected (by one, or more domain experts) in order to pinpoint somewhat reliable class labels for the training patterns. Instances of the difficulties involved in the process are found in areas such as bio-informatics, speech processing, or affective computing, where the exact class labels may not even

be explicitly observable. Although annotating data might be extremely difficult and time consuming (or, sometimes, even impossible), supervised learning is still far the most prominent branch of machine learning and pattern recognition [227].

In archaeology, it is also common to come across settlements that produce pottery of multiple/different kind/type and therefore class. The end composition of a ceramic artefact is highly dependent on its source material and it is therefore noted that classes exhibit a degree of similarity based on the amount of mutual ingredients; in chemical compositional data this relationship is expressed in the concentration of main and trace elements. Additionally, there has been the claim that most “big data” research uses masses of simple and homogeneous data, whereas archaeologists struggle with the variety and complexity of their data sets [99].

Providing homogeneous and consistent access to data might hide the diversity of interpretation that data inherently support. In other words, blindly re-using somebody else’s data could lead researchers to disregard the implicit assumptions embedded in those data and make a bad use of the additional information provided in this way.

5.2.1 Chemical Compositional Data

Chemical analysis is involved in enumerating the number of each type of atoms in a sample; concentrations are usually given in relative numbers (as percentages (%) or as parts per million (ppm)). During an artefact’s analysis, it is common to collect multiple measurements (3 or 5) with the same technique from different positions in an effort to better capture its overall composition. The analyst will then retain measurements, found to be representative, and will then average those to produce a feature vector $x_i \in [0, 1]^p$ where p is the number of analysed chemical elements.

The chemical constituents of any specimen, can be categorised into main and trace elements. Main elements comprise large proportions of the specimen under analysis, while trace elements are present in concentrations less than 0.1%. An example of an artefact’s chemical analysis is shown in Figure 5.1, where the values of pick compositions are extracted and assume the artefact’s composition; the microstructure of the same artefact obtained with SEM (Scanning Electron Microscope) analysis is shown in Figure 5.2.

Additionally, as we deal with heterogeneity in composition data, with the majority of their major elements present in most specimen, the discrimination of objects into

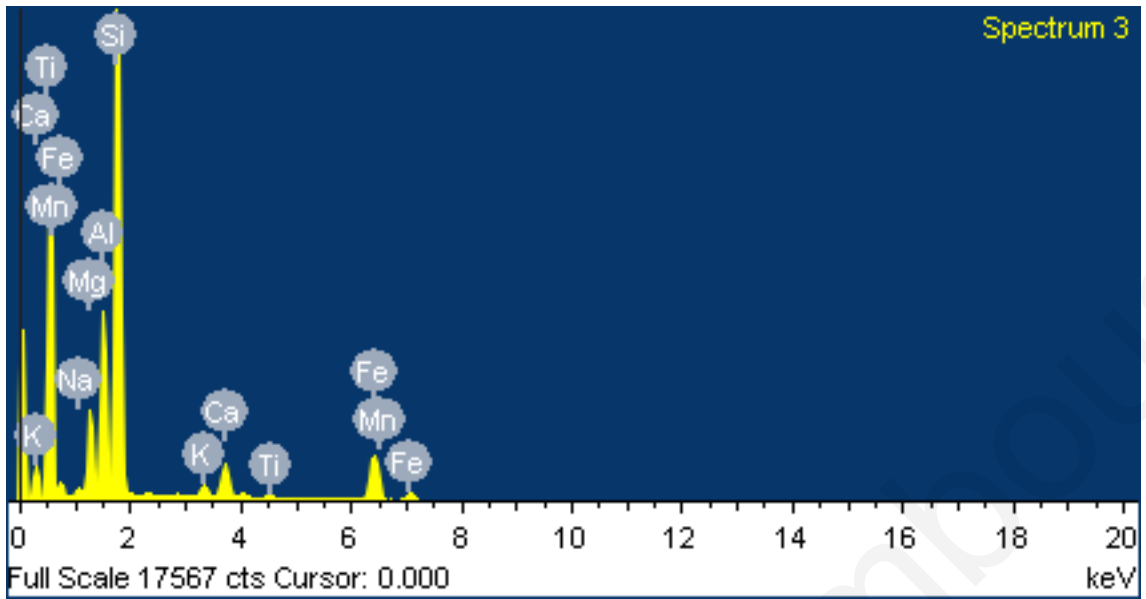


Figure 5.1: Spectrum of Analysed Artefact ©Maria Dikomitou-Eliadou

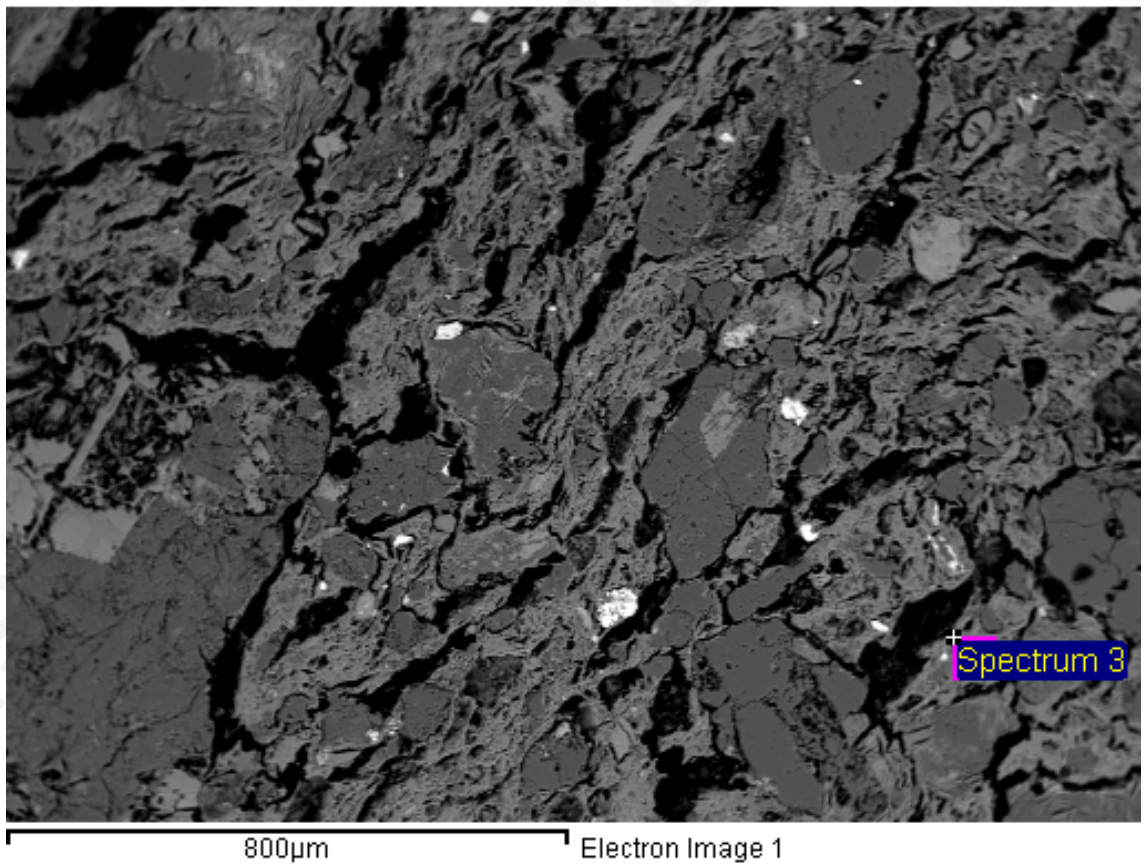


Figure 5.2: Microstructure of Artefact ©Maria Dikomitou-Eliadou

groups makes necessary the utilisation of trace elements in determining the fingerprint of a deposit [174]. As ceramics are heterogeneous in composition with the majority of their major elements present in most artefacts, the discrimination of objects into groups makes necessary the utilization of trace elements.

5.3 Classification in Archaeology

Common analytical questions include those relating to the existence of chemically distinct groups within the data and their association with different manufacturing technologies and their origins [141]. Over the last few decades, the deployment of advanced statistical methods has been proposed in assisting answering various archaeological questions. One of the main concerns in the areas of pattern recognition and data mining is how to organize observed data into meaningful structures. Within the context of archaeology, this can provide answers in the main concerning areas of archaeologists, the distribution and production of artefacts.

The roots of classification analysis of archaeometric data are traced back multiple decades ago with the contribution of Kowalski in 1972 [149] being an early landmark. In subsequent years, classification methods have been used in a number of studies [19,90,148,163,181]. A clear milestone in the analysis of archaeometric data is Baxter's work in 2006 [20], where he reviews the application of classification methods (among others) on the chemical composition of glass artifacts [20]. The effectiveness of a variety of classification methods was evaluated; among them also the three methods to be examined in this chapter for the validation of the methodology. Despite this, the results of the two works are not be straightforwardly comparable due to the different experimental data and deployed methodology.

The complexity and dimensionality of the data makes necessary the use of machine learning methods to allow the characterization of samples based solely on their constitution, while knowing that the presence of noise in the data makes the detection of the clusters even more difficult. The number of studies with primary objective the characterization of specimens is countless and they all share the need to identify some structure or pattern in the ceramic body mainly through chemical analysis. This is all based on the assumption that ceramics made from the same raw material will be similar to a certain degree chemically.

5.3.1 Problem Formulation

Since archaeological data emerge from tangible artifacts, the definition of the classification problem is slightly different. Assuming a set of archaeological artifacts, an operational definition of classification of archaeological chemical data can be stated as follows: given a set of n archaeological artifacts and a vector t indicating the label of each artifact find a model which successfully assigns new samples to the appropriate class.

Consider a set of n archaeological observations $O = o_1, \dots, o_n$. The analysis of the actual-tangible artifact o_i with the use of ED-XRF or any other chemical method of ancient pottery analysis will produce the chemical representation of o_i which has the form $x_i \in [0, 1]^p$ where p is the number of analysed chemical elements and vector x_i consists the chemical compositions of the artefact. Therefore the set $X = \{x_1, \dots, x_n\}$ represents the dataset of the sediments' chemical composition; a set of quantifiable features. We assume that there are groups (subsets) of similar sediments in O , the class of which is determined by the labels in $Y = \{y_1, \dots, y_n\}$, and $y_i \in J$ where J denotes the possible (known) class labels. Each artifact is represented in the dataset with the pair $o_i = (x_i, y_i)$, during parameter training the parameters Θ of the classifier are obtained by $\Theta = g(O)$, and the class t_i of an uncategorised set of artifacts $X = \{x_{n+1}, \dots, x_{n+l}\}$ can be obtained by $y_i = f(x_i, \Theta), \forall n < in + l$.

5.3.2 Analysis Practices for Compositional Data

Chemical compositional data are defined as vectors of strictly positive components, usually expressed as percentages or parts-per million (ppm), with constant sum, a restriction not always maintained. Quantitative chemical analysis is not involved in measuring, but in enumerating or counting the number of each type of atoms in a sample [41]. Chemical compositional data do not vary independently and concentration based approaches to data analysis can lead to misleading conclusions [204].

Chemical compositional data therefore lay in the constrained Simplex Space [6] [41], where correlation analysis and the Euclidean distance are not mathematically meaningful concepts [204]. Furthermore, graphical depiction of raw or log transformed data should only be used in an exploratory data analysis sense, to detect unusual data behaviour or candidate subgroups of samples [2].

Standard multivariate analysis designed for unconstrained multivariate data is en-

tirely inappropriate for the statistical analysis of compositional data. This is due to the fact that the space of compositions is a simplex, a generalisation of a triangle and tetrahedron space, radically different from the real Euclidean space (the space for representing unconstrained vector data).

The chemical constituents of an archaeological artefact, or any other object, can be categorised into major and trace elements. Major elements comprise large proportions of the artefact under analysis, while trace elements are present in concentrations less than 0.01%. As ceramics are heterogeneous in composition with the majority of their major elements present in most artefacts, the discrimination of objects into groups makes necessary the utilisation of trace elements in determining the fingerprint of a deposit [174].

The need for the implementation of methodologies appropriate for deployment in the simplex space has been expressed for over a century, such as: [48, 191, 219, 270]. All pointing out to the conclusion that product-moment correlation of raw components is a meaningless descriptive and analytical tool in the study of compositional variability. Formally the sub-composition based on parts $(1, 2, \dots, C)$ of a D -part composition (x_1, \dots, x_D) is the $(1, 2, \dots, C)$ -sub-composition (s_1, \dots, s_C) defined by

$$(s_1, \dots, s_C) = (x_1, \dots, x_C) / (x_1 + \dots + x_C)$$

As a result any attempt in producing compositional statements for comparison or correlation related purposes with sub-compositions will result in misleading results. Ignoring the principle of sub-compositional coherence has been a source of great confusion in compositional data analysis [7].

5.3.3 Data Analysis in the Simplex

Chemical compositional data (see Figure 5.3) lay in the constrained Simplex Space [6, 41], where correlation analysis and the Euclidean distance are not mathematically meaningful concepts [204]. Formally, compositional or closed data refer to p – *dimensional* vectors $\mathbf{x} = [x_1, x_2, \dots, x_p]$ of positive components summing up to a constant k , hence defined on the simplex sample space: $S^p = \mathbf{x} = [x_1, x_2, \dots, x_p] | x_i > 0; \sum_{i=1}^p x_i = k$ [176].

The characterisation of S^p as Euclidean space related topics are summarized in [176] and [8]. From the Euclidean structure of S^D , a distance, known as the Aitchinson

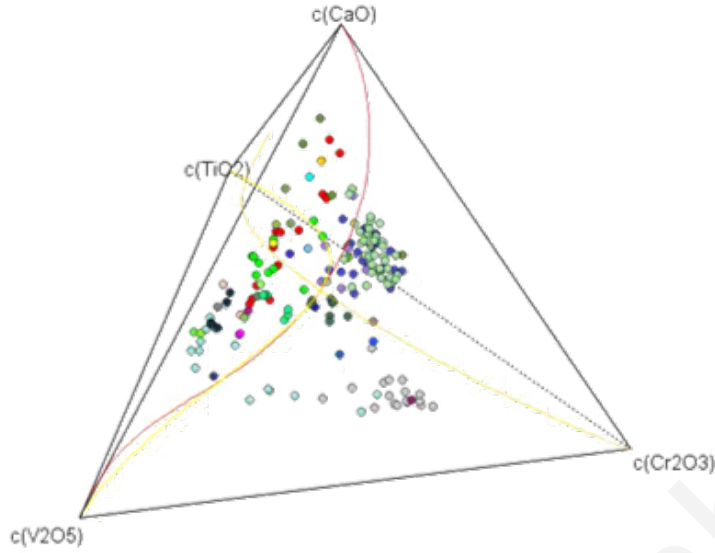


Figure 5.3: Representation of data in the constrained simplex space

distance [5] is included:

$$d_a(\mathbf{x}, \mathbf{x}') = \left[\frac{1}{D} \sum_{i < j} \left(\ln \frac{x_i}{x_j} - \ln \frac{x'_i}{x'_j} \right)^2 \right]^{\frac{1}{2}} \quad (5.1)$$

The characteristics of the Aitchison distance are: Scale invariant, Permutation invariant and Sub-compositional dominant. The most important characteristic of compositional data is that they carry only relative information and therefore scale invariance is related to universality as features should not change if compositions are multiplied by a common factor. Additionally, a function is permutation-invariant if it yields equivalent results when we change the ordering of our parts in the composition. The final condition is sub-compositional coherence: sub-compositions should behave as orthogonal projections do in conventional real analysis. The size of a projected segment is less or equal than the size of the segment itself. This general principle, though shortly stated, has several practical implications, the most illustrative are that the distance measured between two full compositions must be greater (or at least equal) than the distance between them when considering any sub-composition and that if we erase a non-informative part, our results should not change.

When the analysis involves data that row-wise do not sum up to 100 another column is added forcing the sum of each row to be at 100. For example if the sum of n elements is 98 then the cell in the added column (the $n^{\text{th}} + 1$) will have the value of 2. A sub-compositionally dominant distance implies that the distance measured in the $n+1$ parts must be greater or equal to the distance measured in the n part sub-composition. The

use of the Aitchison distance allows the deployment of classification methods without violating restrictions of the Simplex.

Another option in utilising ML methods on compositional data is their transformation to the Euclidean unconstrained space. This is possible through the use of log-ratio data transformations (additive, centered and isometric). Following [6], the statistical analysis of compositions should be focused on the ratios between components. In this way, data are moved to real space wherein standard methods can be applied. Log-ratio data analysis relies on the fact that in high dimensional problems (many features) it is often desirable to select the most informative features and therefore acts in a sense as a dimensionality reduction. This is done in an effort of problem simplification for achieving possibly better generalisation, however this approach diminishes the contribution of minor elements and therefore it does not serve as a possible option in our analyses.

Nonetheless, any pottery analysis is not a straightforward process, and there are various parameters (i.e. contextual, spatial, chronological, compositional, technological) that the researchers need to consider while defining their research design, their sampling strategy, and later while evaluating their research results.

5.4 Validation of Methodology on Ceramic Data

The methodology introduced in Chapter 4 has been applied on a sample of utilitarian pottery, characterised as scarce with uncertainty that could not allow the expert archaeologist to categorise all samples to a class. In this experiment, the aim was rather to differentiate the specimen based on their fabric, and therefore investigate the degree of similarity between types than to achieve perfect classification or to discriminate the origin of each artefact.

The results of this experiment are originally published in [46] with analysis on a sample of Early and Middle Bronze Age utilitarian pottery from Cyprus. The statistical experiment involved two analytical datasets deriving from the mineralogical and chemical characterisation of 177 ceramic samples, with the respective employment of ceramic petrography and ED-XRF [71]. The samples were categorised in 15 groups; 22 samples were marked as outliers due to their condition and are usually neglected in further analysis. Data annotation (labeling) has been performed based on the artefacts' microstructure, macroscopic attribute and petrographic analysis.

The objective of the experiment was to evaluate the algorithms' performance – in terms of some cluster validity index – on the given data and validate the result's significance against the shaped hypothesis. For demonstration purposes and for the needs of this experiment, classification is achieved with three well-known methods, a standard statistical method “k-Nearest Neighbours” (k-NN) [76], a method using “Decision Trees” (C4.5) [201] and a more complex structure with foundations in neural networks “Learning Vector Quantisation” (LVQ) [145]. The selection of these three algorithms was driven by the need to test the effectiveness of different types of algorithms on the analysis of archaeological data, however, the presented design may be realized with any classification method.

5.4.1 The Archaeological Dataset

This statistical study involves the compositional analysis of a small elemental dataset obtained from the ED-XRF analysis of Early and Middle Bronze Age ceramics from Cyprus [71]. The archaeological samples derived from the occupational phases of the Early and Middle Bronze Age settlement of Marki Alonia in central Cyprus, and include the two predominant wares recorded at the site, i.e. Red Polished Philia pottery from the first occupational phases of the settlement and Red Polished pottery from the Early and Middle Bronze Age. Red Polished Philia pottery was also selected from other contemporary sites across Cyprus, in order to assess the degree of compositional and technological homogeneity among pottery assemblages that exhibit a significant degree of stylistic uniformity across the island. Therefore the final 177 samples under study come from eight different sites across the island [71]. Their analytical study aimed at their compositional and technological characterisation in order to assess ceramic production, distribution and social interaction in Early and Middle Bronze Age Cyprus [71–73].

The samples were divided into two datasets. The first dataset involved the Red Polished Philia samples from various Philia sites, while the second dataset involved all the samples from Marki Alonia, including both the Red Polished Philia and Red Polished samples from the settlement. The statistical experiment was particularly challenging due to its small size (177 samples), and the relatively large number of outliers (21 samples), which were not categorised – by the expert – in one of the predefined groups, either because they lack discriminating petrology, or because their

fabric was dissimilar to those of the clustered samples.

At this point, it is worth noting that the labelling procedure was performed by the expert archaeologist utilising knowledge other than the chemical compositions of the samples (i.e. petrography). The samples were labelled into 15 fabric groups (Marki fabrics I-XIII and Philia fabrics I and IV). Considering the fact that the identification of outliers is as important for the assessment of ceramic compositional and technological variability, all outliers were also included in the experiment, each outlier forming a separate class, resulting in a total of 36 different classes. The consideration of outliers served a twofold purpose:

1. To test the robustness of classification on complicated and highly overlapping data, and
2. To assess whether post-classification analysis could allow outlier categorisation to one of the predefined fabric groups (a task that could not be solved with certainty beforehand due to the absence of discriminating petrology in the ceramic thin sections).

Finally, the statistical experiment was conducted in order to explore other methods of statistical analysis that are not yet widely known in the field of archaeological sciences and investigate the relations among fabric groups within the two datasets that are suggested by petrography to be identical or very similar, with the ultimate objective to test the correspondence between the mineralogical and chemical compositional data.

The dataset became subject to treating before statistical analysis. All elements were converted into oxide compounds with stoichiometry, the composition of each artefact was normalised to allow the application of Aitchinson distance (i.e. force the sum of each row to be 100). Trace elements with elemental concentration below 10 ppm were omitted along with sulphur trioxide (SO_3), chlorine (ClO) and lead oxide (PbO) concentrations due to analytical reasons. It is a typical practice in chemical compositional analysis to exclude features with very small concentrations due to instruments inability in accurately enumerate in very low ranges. Sodium oxide (Na_2O), phosphorus pentoxide (P_2O_5), cobalt (Co_3O_4) and cerium oxides (CeO_2) were also omitted from multivariate statistics due to inconsistencies in values and poor reproducibility in successive analytical runs [71]. The chemical compounds used for analysis are: MgO , Al_2O_3 , SiO_2 , K_2O , CaO , TiO_2 , V_2O_5 , Cr_2O_3 , MnO , Fe_2O_3 , NiO , CuO , ZnO , Ga_2O_3 , Rb_2O , SrO , Y_2O_3 , ZrO_2 , BaO .

5.4.2 Hypothesis Testing

It should be clarified that the two different datasets, i.e. the “Marki” and “Philia” datasets, were studied in the same period of time and by the same researcher, therefore they are characterised by a reliable degree of consistency, as all samples were at first collected and then analysed by the same person in the same laboratory, using the exact same procedures. The combination of the data into a single dataset could, however, reveal compositional and/or technological relationships between types/classes of ceramics, as well as links emerging either due to their context of production or recovery and/or their technology of production.

The null hypothesis behind the classification problem stated that the classification algorithms, k-Nearest Neighbour, C4.5 (based on Decision Trees) and Learning Vector Quantisation (LVQ Networks) perform equally well for the dataset of interest when the performance is measured with the classification accuracy. Upon the rejection of the null hypothesis, the alternative hypothesis tests whether any of the three algorithms outperforms the other in pairwise comparisons. Each of the three selected algorithms operates on different principles, each representing a different category of statistical/machine learning approaches. Doing so allows solving the classification problem from different perspectives, hoping that the differences in the results may disclose new information about the data.

5.4.3 The experimental design

The proposed methodology is subsequently used to differentiate a series of ceramic specimens based on their fabric, and investigate the degree of similarity between discriminated types. The choice of deployed methods accounts for the non-uniformity in the distribution of classes, as well as the size of the dataset, which is relatively small compared to the number of classes.

Classification of small datasets (with regards to the number of classes) requires the use of re-sampling methods. For the needs of this study, bootstrapping with replacement is used to allow the generation of datasets of 177 samples. The choice of re-sampling adds the assumption that the dataset is representative to the population, in other words, the sample of collected ceramics depicts a good representation of the ceramic population at the sampled sites, during the specified time period. Moreover, the performance of the algorithms is evaluated against the classification accuracy; the

number of correct predictions from all predictions [232]. The calculation of the accuracy for a problem with more than two classes also accounts the instances in which the classifier correctly decides not to assign an artefact to those fabrics, to which it does not belong. The classification result is also evaluated against the Jaccard index — an external cluster validity index — which is calculated as the number of correctly classified samples over the number of samples that exist either in the true or estimated classification [240]; this index does not account for instances in which the algorithm correctly did not assign a sample to a specific class. The calculation of both indices is done in an effort to observe their robustness and emphasise the significance of their choice. It is not expected that an algorithm will perform equally well with any dataset. The results are highly dependent on the parameterisation of the algorithm, as well as the structure and complexity of the data. Since the evaluation of classification performance forms part of the question, it was necessary to consider some fine-tuning of each algorithm's parameters for each bootstrapped dataset. However, the fine-tuning of each algorithm's parameters cannot be assumed to be equivalent to one another. For instance, the LVQ algorithm requires the parameterisation of more parameters than the other two and its performance is more likely to suffer.

k- Nearest Neighbour (k-NN)

The k-NN algorithm is a non-parametric approach used for classification, operating in the belief that a sample will more likely belong to the class of its closest already classified artefacts [76]. k-NN is among the simplest and most intuitive machine learning algorithms. For each uncategorised artefact, its distance to all classified samples is measured, the k closest samples are selected and the artefact is categorised to the class most of the k neighbours belong; k is a positive integer, typically small. If $k = 1$, then the object is simply assigned to the class of its immediate nearest neighbour, according to some distance metric. The input consists of the k closest training examples in the feature space and the output is a class membership. The valid implementation of the algorithm for compositional data in the Simplex space requires measuring the distance with the Aitchison distance metric [6].

C4.5 Algorithm

The C4.5 is an extension of the earlier ID3 algorithm and performs classification by generating a decision tree [152]. Decision trees are generated incrementally by breaking down a dataset into smaller and smaller subsets. C4.5 builds decision trees from a set of training data, using the concept of information entropy. At each tree node, the chemical element/feature that most effectively splits the dataset into subsets is selected while the attribute with the highest normalized information gain is chosen to make the decision. The idea is to refine T (the tree) into subsets of samples that are heading towards single-class collections of samples. An appropriate test is chosen, based on a single element that has one or more mutually exclusive outcomes [248]. The decision tree for T consists of a decision node identifying the test and one branch for each possible outcome (see Figure 5.4). The C4.5 algorithm then recurs on the smaller sub-lists. The decision trees generated by C4.5 can be used for classification and it is referred as a statistical classifier.

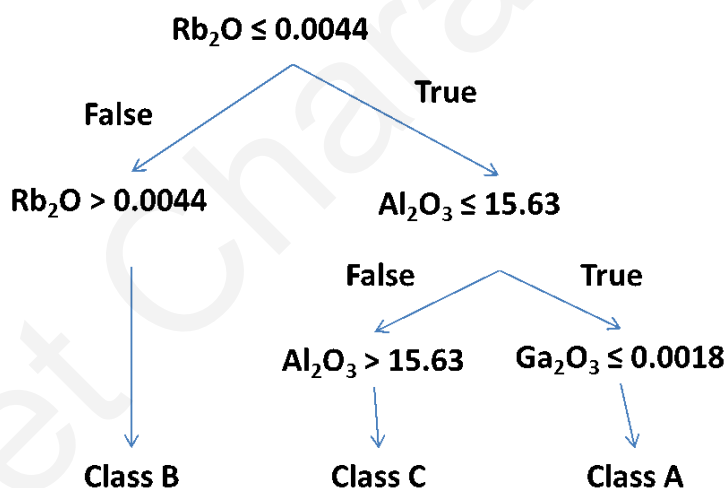


Figure 5.4: Decision Tree example based on the values of three chemical elements

Learning Vector Quantisation (LVQ)

Learning Vector Quantisation is a special case of an artificial neural network which deploys the winner takes it all learning-based approach. An LVQ system is represented by prototypes $W = w_1, \dots, w_m$ where m is the number of classes defined in the feature space of observed data. Algorithms based on this approach, assign each data sample, the label of the prototype that is closest to it, according to a given distance measure. The position of this so-called winner prototype is then adapted, i.e. the winner is

moved closer if it correctly classifies the data point or moved away if it classifies the data point incorrectly. An advantage of LVQ is that it creates prototypes that are easy to interpret. A trained LVQ network allows the visualisation of the map of prototypes which gives insight on which classes are closer to others. In archaeological context this may provide information as of how “close” fabrics are to each other. Even though the algorithm does not restrict the dimensionality of the map, it is usually implemented as a 2D map.

Algorithm Parameter Estimation

The deployment of the discussed algorithms for each bootstrapped dataset required some fine-tuning of the algorithms. However, exhaustive fine-tuning might easily become very time consuming especially when a large number of re-sampled datasets need to be processed. The set of categorised samples is used for the parameterisation of the algorithm and it is further divided into 2 smaller sets named the training tuning and testing tuning sets. The parameters maximising the classification performance of the algorithm on the testing set are selected for processing the bootstrapped dataset. Considering the very restrictive size of the training and testing tuning datasets, it is expected that the selected parameters might not be the optimal; a decision which we are obliged to make to avoid violating the rules of training classification algorithms [76]. The C4.5 algorithm did not become subject to parameterisation or pruning. Pruning is a way of reducing the size of the decision tree, for this experiment, the parameter determining the pruning stage was set to 0 corresponding to no pruning. The k-NN method only requires the specification of the parameter k . For each dataset the algorithm was tested for the integer values of $k = 1, \dots, 10$; 96% of the time the values of k maximizing the classification accuracy were between 1 and 4, with $k = 1$ being the most frequently occurring value scoring 73%. LVQ required most of the training; the configuration of the network requires specifying the learning rate and map size in each dimension; for convenience this study was limited to 2 dimensional square maps —the most common implementation [145]. The tuning of an LVQ network has a significant computational cost, and due to this the parameterisation was limited to a small range of possible values.

5.4.4 Statistical Testing

The 5×2 cross validation paired t-test [68] and the 5×2 cross validation F-test [10] were deployed to firstly statistically test the significance of the classification results and secondly to evaluate their robustness. Benchmarks on significance testing propose that cross validation testing methods are more robust when dealing with small datasets where reproducibility of the experiment is not an issue [68]. The 5×2 cross validation methods were selected to allow large enough datasets for testing while ensuring that no further dependencies of overlapping training and testing sets are introduced when cross validation is used [217].

The experiment was allowed to run for 500 iterations to allow the generation of valid statistics and the significance of the results was calculated at level 0.05 ensuring that there is 95% confidence that the results of statistical testing represents the reality.

5.4.5 Results and Discussion

Figure 5.6 shows a plot obtained by Linear Discriminant Analysis (a dimensionality reduction method) [29] of the original dataset. Many classes are overlapping and the discrimination of classes is not trivial. Table 5.1 shows the performance, in terms of classification accuracy and the Jaccard index. The scores are calculated as the mean scores of all iterations. As expected the values scored measuring the classification accuracy are higher than the values scored by the Jaccard index. In archaeological ancient pottery analysis, it is important to measure a classifier's robustness in assigning artefacts to the correct fabric. Since classification accuracy accounts the instances in which the classifier correctly does not assign an artefact to fabrics to which it does not belong it served better the needs of the problem compared to the Jaccard coefficient. The classification accuracy scores for the three algorithms is summarised in Figure 5.5. The classification evaluation scores are seemingly low. This is due to the high overlapping of possible classes and their under representation. Additionally, as previously stated, the classification accuracy metric does not account for robustness in the separability of the data.

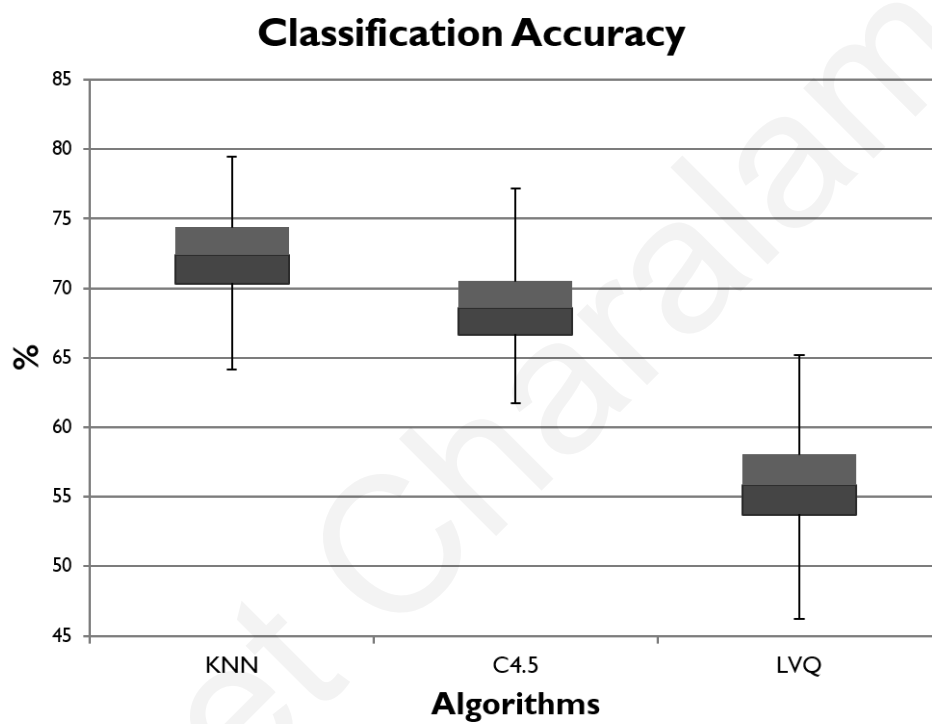


Figure 5.5: Variability in classification accuracy between algorithms

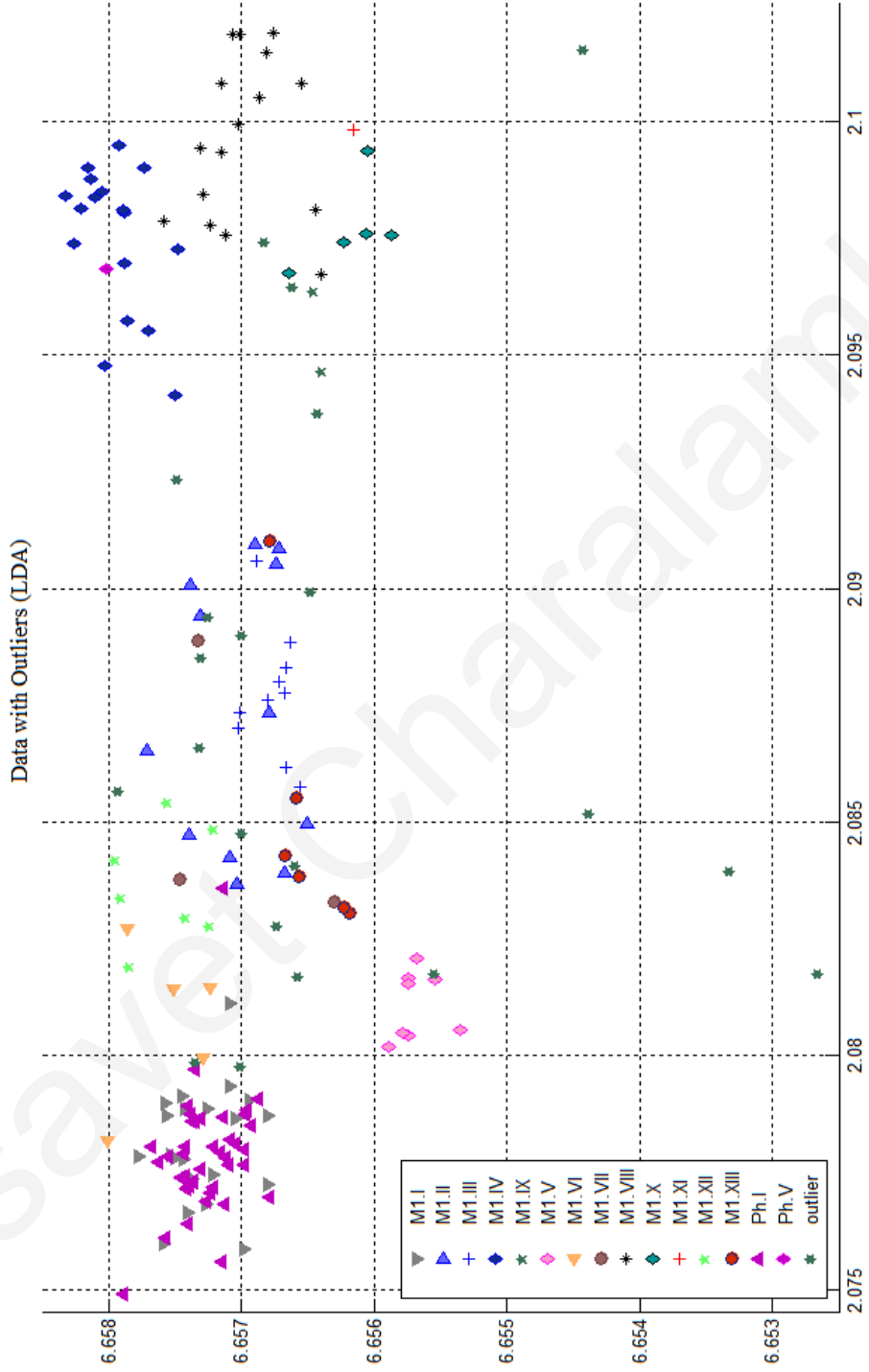


Figure 5.6: Two dimensional plot of the sample generated with LDA. The fabrics are not distinct in the state space, many classes overlap

The scores claim that the k-NN algorithm consistently scores the highest compared to the other algorithms. Significance testing between the algorithms, in a pairwise fashion, has illustrated that the performance of k-NN and C4.5 outperformed the LVQ Network, while the Null hypothesis between the k-NN and C4.5 (stating their non significant difference performance) is accepted. In other words, k-NN and C4.5 score better than LVQ, however, k-NN is not significantly better than C4.5. It is also important to note that both statistical tests, the 5x2 cv paired t-test and the 5×2 cv F-test, confirmed the same hypothesis results.

Table 5.1: The estimated accuracy and Jaccard index of each algorithm. The scores represent the mean score of all iterations.

Algorithm	Classification Accuracy (%)			Jaccard Index (%)		
	Mean	Max	Min	Mean	Max	Min
<i>k - NN</i>	72.1	79.4	64.2	56.7	70.1	42.7
C4.5	68.5	77.2	61.7	49.1	63.7	38
<i>LVQ</i>	55.8	65.2	46.2	30.3	38.8	21

The experiment does not show that LVQ performs worse; it is rather shown that its operation with the very limited fine-tuning and the discussed dataset, results in lower performance compared to the other classification methods. LVQ is admittedly a more complex algorithm and its parameterisation needs to be handled with care, especially when dealing with datasets of very limited size containing a large number of classes. The, potentially, poor selection of parameters, during fine-tuning, may hinder the classification results; something that became apparent in the performance of the LVQ. When one deploys any classification approach should really understand the operation of the algorithm and its appropriate configuration.

The results of classification can be useful in verifying the initial distinction of the samples into fabric groups and outliers. Most importantly, further analysis on the classification results has shown that classification may provide more information to the archaeologist. During each iteration, a matrix with the correctly and wrongly classified artefacts for each class (i.e. the confusion matrix) is generated. Systematic study of the misclassified artefacts has shown that some elements are not misclassified randomly. For instance, elements of class M1.II if they were to be misclassified then the algorithm would assign them to class M1.III; the same holds for all classes shown in Table 2.

Indeed the comparative study of the fabric groups suggested by petrography and the highlighted elemental relations suggested by this statistical experiments confirmed that fabrics M1.I and Ph.I are essentially the same fabric distributed across Cyprus during the Philia cultural phase, including the settlement at Marki, while M1.II and M1.III share many mineralogical characteristics, being made with raw materials deriving from a similar geological environment. It was also very interesting to see that the algorithms proposed a link between the two most igneous fabrics, those being M1.VIII.

Table 5.2: Inter-class relationships in a multi-class problem.

<i>Class1</i>	<i>Class2</i>	<i>SampleID</i>
M1.I	Ph.I	
M1.II	M1.III	
M1.II	M1.VII	
M1.II	M1. outlier10	14604
M1.III	M1.II	
M1.IV	M1.VI	
M1.IV	M1. outlier17	16513
M1.V	M1.XIII	
M1.X	M1.VIII	
M1.XII	M1.VII	
M1.XII	M1. outlier6	12372
Ph.I	M1.I	

Another important finding is that the analysis of misclassified artefacts suggested a possible class (fabric group) to some outlier samples. Inter-class relationships within the context of archaeology are plausible as different pottery fabrics might be produced in the same settlement, with the use of mutual ingredients or even present evolution changes. This analysis becomes useful for the classification of samples of unknown class with a degree of confidence (analogous to the classification accuracy). The result of this process returns a possible class for each of these samples which when interpreted may lead to their definite categorisation. The inability of the algorithm to classify the specimen to their true class reveals possible relationships between certain classes (see Table 5.3). Having said that, it is not anticipated that all misclassified artefacts admit correlations between classes; it is expected the classifiers to adhere to certain error

rates.

Table 5.3: Classification accuracy and Jaccard index scores when classification is performed on elements with mean concentration $>0.1\%$ and $<0.1\%$.

		Classification Accuracy (%)			Jaccard Index (%)		
	Elements Used	Mean	Max	Min	Mean	Max	Min
$k - NN$	$>0.1\%$	73.2	79.5	67.6	52.9	64.8	40.5
	$<0.1\%$	66.6	75	58.2	47.8	57.7	35.7
C4.5	$>0.1\%$	67.8	75.1	62.7	43.2	56.8	32.8
	$<0.1\%$	64.5	71.2	55.8	46.1	58.1	35.6
LVQ	$>0.1\%$	57.3	67	48.8	29.3	36.6	20.9
	$<0.1\%$	59.1	66.2	51.4	40.2	52.2	26.1

The specimens of request were analysed for a number of chemical elements, some of which in very small concentrations ($< 0.1\%$). The heterogeneous composition of ceramics needs to be accounted during classification. Trace elements may concur more characteristically in determining the fingerprint of a deposit [173], making important the evaluation of their discriminating abilities. Due to this, the experiment as discussed previously was repeated two more times, using the chemical elements with mean concentration $> 0.1\%$ (MgO , Al_2O_3 , SiO_2 , K_2O , CaO , TiO_2 , MnO , Fe_2O_3 , BaO) and another one using only the chemical elements with mean concentration $< 0.1\%$ (V_2O_5 , Cr_2O_3 , NiO , CuO , ZnO , Ga_2O_3 , Rb_2O , SrO , Y_2O_3 , ZrO_2); in both cases the data rows were normalised to sum 100. The results of the experiments bring to our attention some interesting data properties summarised in Table 5.2 where each table column shows potential relationship between Class 1 and Class2. Upon classification, samples of Class 1 in case they are misclassified, they would more likely be allocated to Class 2. The exclusive use of elements with mean concentration $>0.1\%$ allows the equivalent discrimination of artefacts when using all available information (see Table 5.1). Table 5.3 also shows that despite not utilising 99.8% of the measured information (when $<0.1\%$), the majority of characteristics that allow the discrimination of the specimen into their categories is maintained. This finding allows us to hypothesise that the use of trace elements during classification needs to be studied further.

Chapter 6

Case study II: Acoustic Event Detection

6.1 Introduction

Acoustic event detection and classification (AED/C aka AED) is a recent discipline that may be included in the broad area of computational auditory scene analysis. It consists of processing acoustic signals and converting them into symbolic descriptions corresponding to a listener's perception of the different sound events present in the signals and their sources. AED aims to identify both timestamps and types of events in an audio stream. This becomes very challenging when going beyond restricted highlight events and well controlled recordings. AED escapes the field of speech recognition as it is concerned about both voiced and unvoiced auditory scenes. Depending on the approach, it is often observed that machine learning methods through classification and clustering techniques come into play. Additionally, detection and classification of sounds other than speech may be useful to enhance the robustness of speech technologies like automatic speech recognition.

The auditory scene in an environment is highly dynamic – and unpredictable – making the training of robust classifiers a real challenge. Acoustic events may adhere to very diverse characteristics both in terms of duration and frequency content. The perception of sound is significantly impacted by the surrounding environment and it is enhanced with reverberations, created as a result of sound reflections on adjacent objects and the environment; the duration and impact of reverberations are dependent on the signal's frequency. As a result, numerous reflections of the sound wave build up and then decay as the sound is absorbed by the surfaces of objects in the space. This is most noticeable when the sound source stops but the reflections continue, decreasing

in amplitude, until they reach zero amplitude.

The accurate detection and recognition of acoustic events in the highly dynamic auditory environment require large enough datasets to allow appropriate representation of the event under different conditions. Several hours of the recorded audio of the same event are required. Considering that the duration of acoustic events varies drastically according to its type, it is not always possible to generate sufficient data for training. Data scarcity and class under-representation impact greatly the model training process. As a result, often researchers are forced to train classification models utilising data of different – and heterogeneous – encoding schemes; introducing further uncertainty and noise in the extracted features.

Acoustic events such as explosions and gunshots in particular, which are of primary interest in the field of AED, are heavily impacted by reverberations as they are characterised by a short (in time) impulse of high amplitude. The perception of such events by the human ear differs significantly, as the impulse responses of the same event in closed and open environments are drastically different.

It is a common practice in AED to split the audio stream in chunks for feature extraction and subsequent processing and it is also very common to use frequency domain characteristics as features. The isolation of short-time audio chunks makes its recognition and categorisation even more challenging when put out of context; blocks are processed in isolation without memory. Additionally, events of interest in AED, range in characteristics and may be both voiced and unvoiced. The production of ML models capable of robustly identifying events need to involve the use of appropriate features to allow for this; it is also critical to cope with signal continuity in the case of streamlined data.

The existence and timestamps of many non-speech sounds, i.e. (non-speech) acoustic events, reveal human and social activities. Such information is very helpful in applications such as surveillance, multimedia information retrieval and intelligent conference rooms. Additionally, efforts have been made to produce unsupervised clustering of interesting events recorded automatically in an office environment [111] to allow the extraction of highlights in the audio stream.

In the context of surveillance systems, AED aims to fill the gap when other analytics are not in the position to provide results in dark environments, incidents residing outside the field of view or the inability to detect events due to overcrowded areas.

6.2 Audio Coding

In digitization, a microphone detects changes in air pressure and sends corresponding voltage changes down a wire to an ADC (Analog to Digital Converter) which regularly samples the values. In the realm of sound, the digitization process takes an analog occurrence of sound, records it as a sequence of discrete events, and encodes it in the binary language of computers. Digitization involves two main steps, sampling and quantization.

Sampling is a matter of measuring air pressure amplitude at equally-spaced moments in time, where each measurement constitutes a sample. The number of samples taken per second (samples/s) is the sampling rate. Units of samples/s are also referred to as Hertz (Hz). The Hertz unit is also used to mean cycles/s with regard to a frequency component of sound.

Quantization is a matter of representing the amplitude of individual samples as integers expressed in binary. The fact that integers are used forces the samples to be measured in a finite number of discrete levels, whose range is determined by the bit depth (the number of bits used per sample). A sample's amplitude must be rounded to the nearest of the allowable discrete levels, which introduces error in the digitization process. An example of this process is shown in Figure 6.1.

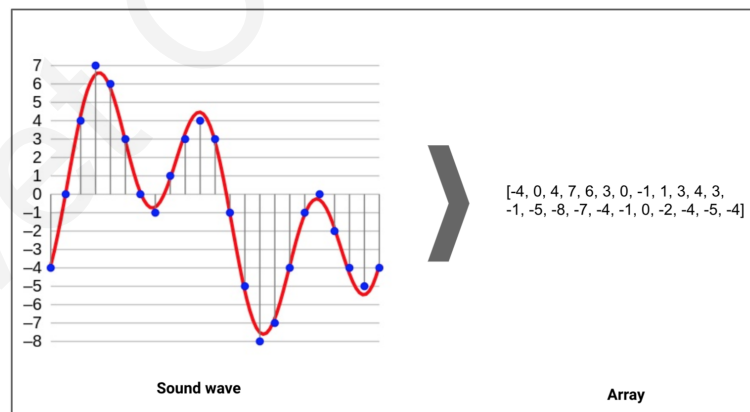


Figure 6.1: A sampled and quantised sound wave

A sound wave, in red, represented digitally, in blue (after sampling and 4-bit quantisation), with the resulting array shown on the right.¹

The above image shows how a sound excerpt is taken from a waveform and turned into a one dimensional array or vector of amplitude values. The samples are stored as

¹Source:<https://commons.wikimedia.org/wiki/File:4-bit-linear-PCM.svg>

binary numbers. From these stored values, the amplitude of the digitized sound can be recreated and turned into analog voltage changes by the DAC.

The quantity of these stored values that exist within a given amount of time, as defined by the sampling rate, is important to capturing and recreating the frequency content of the audio signal. The higher the frequency content of the audio signal, the more samples per second (higher sampling rate) are needed to accurately represent it in the digital domain. Based on the Nyquist Theorem, and in order to be able to successfully digitize a sound wave, the sampling rate must be at least twice the frequency of the highest frequency component; in any other case an undesirable effect called aliasing is introduced resulting in lowering the frequency content of the signal upon reconstruction.

The described procedure is known as PCM (Pulse Code Modulation) encoding and when the Nyquist theorem is respected the end result is lossless; a perfect DAC (Digital to Analog) reconstruction of the signal is possible. PCM encoded sound waves usually bear the WAV extensions (WAVE codec). Lossless encoding is more expensive in terms of both transmission and storage. Due to this, lossy compression is usually applied on the sampled signal to lessen the dynamic range between the loudest and quietest parts of an audio signal. This is usually done by boosting the quieter signals and attenuating the louder signals. Many compression algorithms rely on the psychoacoustic model (how humans perceive and reconstruct sound) and the audio masking effect (which frequencies are audible in a stream). Therefore, inaudible or close to inaudible frequencies are removed to reduce the signal's bit depth, transmission time and storage size. When lossy compression is applied on a signal, its perfect reconstruction is no longer possible, introducing further errors in the signal. Properties such as insufficient sampling rate and lossy compression (or the use of different codecs) result in the introduction of uncertainty in the signal; also depicted in the feature extraction process.

6.3 Classification in Acoustic Event Detection

6.3.1 Problem Formulation

Most digital signals are infinite, or sufficiently large that the dataset cannot be manipulated as a whole. This is also the case for audio signals. Long signals are difficult to

analyse statistically, because statistical calculations require all points to be available for analysis. To tackle this, the audio stream is divided into blocks (signal subsets) through windowing.

The windowing process will alter the spectral properties of the dataset. A finite window with transfer function $W(z)$ is applied on the digitised signal $S(z)$ as such: $S(\hat{z}) = S(z)W(z)$. The simplest approach of windowing is with the use of a rectangular window, $W(z) = 1$, where all data points before and after the window are truncated.

Parameters such as the length of the window, the window function and the shifting method impact the representation of the signal as distinct blocks of fixed size. The number of blocks depends on the length of the stream and the window length. An operational definition of classification of audio in acoustic event detection can be stated as follows: given a set of n blocks and a vector t indicating the label of each block find a model which successfully assigns new samples to the appropriate class.

Consider a set of n extracted blocks from a continuous stream $O = o_1, \dots, o_n$. Each block, o_i , contains the time-domain quantised values of the audio stream. Each observation undergoes through a feature extraction process to produce a representation of o_i which has the form $x_i \in [0, 1]^p$ where p is the number of extracted features. Therefore the set $X = x_1, \dots, x_n$ depicts the representation of the audio stream as a set of distinct quantitative features. We assume that there are groups (subsets) of similar events in O , the class of which is determined by the labels in $T = t_1, \dots, t_n$, and $t_i \in J$ where J denotes the possible (known) class labels. Each block is represented in the dataset with the pair $o_i = (x_i, t_i)$, during parameter training the parameters Θ of the classifier are obtained by $\Theta = g(O)$, and the class t_i of an uncategorised set of blocks $X = x_{n+1}, \dots, x_{n+l}$ can be obtained by $t_i = f(x_i, \Theta) \forall n < in + l$.

6.3.2 Practices in Acoustic Event Detection

Much research in audio content analysis has typically addressed the problem of segregating a few audio sources [39, 84] or segmenting an audio stream into a small number of acoustically compact categories [197, 197]. Acoustic Event Detection (AED) aims to detect specified acoustic events such as gunshots [56], explosions [43, 182], speech/music transitions [197], cough events [231], or audience cheering at a sports event [15].

Despite the research done so far, reliable detection and categorization of audio events from everyday audio is not mature enough for practical applications and em-

phasizes mostly in monophonic sound. Most of the previous work classifies an audio signal into one of predefined classes using standard features such as mel-frequency cepstral coefficients (MFCC) and classifiers such as hidden Markov models (HMM), Gaussian mixture models (GMM) or other statistical learning methods. Research has also been made on the use of the appropriate features and classifiers for different events. However, comparative analysis on the performance of classifiers on specific events cannot be generalised to cover for all types of acoustic events. One of these studies is implemented for 24 everyday contexts, such as restaurant, car, library, and office in [85]. The system used MFCCs and their first-order time derivatives as features and HMMs with discriminative training for classification. The authors also conducted a listening test to compare the system's performance to the human abilities. The average recognition accuracy of the system was 58%, against 69% obtained in the listening tests, in recognizing between 24 everyday contexts. The accuracy in recognizing six high-level classes was reported as 82% for the system and 88% for the humans. Additionally, [282] proposed extracting discriminative features for AED using a boosting approach, which outperformed classical speech perceptual features, such as Mel-frequency Cepstral Coefficients (MFCCs) and log frequency filterbank parameters with the use of cascaded statistical models and noise adaptive kernels. In particular, a tandem connectionist-HMM approach used to combine the sequence modeling capabilities of the HMM with the high-accuracy context-dependent discriminative capabilities of an artificial neural network trained using the minimum cross entropy criterion. Then, an SVM-GMM-supervector approach was followed implementing noise adaptive kernels better approximating the KL divergence between feature distributions in different audio segments. Experiments on the CLEAR 2007 AED Evaluation set-up demonstrated over 45% relative performance improvement, on detection of twelve general acoustic events in a real seminar environment.

While most of the work in acoustic event detection focuses on a few highlight acoustic events, the 2007 AED Evaluation sponsored by the project "Classification of Events, Activities and Relationships (CLEAR)" [243, 245] was performed on a continuous audio database recorded in real seminars [244]. Systems attempted to identify both the temporal boundaries and labels of twelve acoustic events containing also acoustically subtle or mixed with speech events. The AED evaluation work sponsored by the CLEAR project served as the baseline for other research works. Efforts on acoustic events detection presented in the CHIL project made use of CLEAR AED evaluation [237]. The

goal in the CHIL project was to detect and recognize a closed set of pre-defined acoustic events. The evaluation data consisted of overlapping acoustic events occurring in the CHIL lecture and meeting corpus. Participants to the CLEAR evaluation proposed 5 systems based on HMMs and one on SVMs, using MFCCs and the Viterbi algorithm; the best performing system used HMMs and AdaBoost for feature selection [117, 281].

Considering the success of HMM in the discrimination between acoustic events, efforts have been made to implement systems for acoustic event detection in recordings from real life environments [170], where events were modeled using a network of hidden Markov models; their size and topology was chosen based on a study of isolated events recognition. On real life recordings, the recognition of isolated sound events and event detection was tested. For event detection, the system performed recognition and temporal positioning of a sequence of events. The classifier's performance was measured based on the accuracy metric with a 5-fold cross validation with non overlapping training and testing sets. An accuracy of just 24% was reported in classifying isolated sound events into 61 classes. This corresponds to the accuracy of classifying between 61 events when mixed with ambient background noise at 0dB signal-to-noise ratio. In event detection, the system was capable of recognizing almost one third of the events, and the temporal positioning of the events was not correct for 84% of the time in polyphonic long recordings.

Additionally, efforts have been made to produce unsupervised clustering of interesting events recorded automatically in an office environment [111]. The "interesting" events are detected by continuous monitoring of background noise and then clustered into discrete categories using unsupervised k-means. Authors of [42] propose a framework for detection of key audio effects in a continuous stream with the use of 10 audio effects, distinct enough to be perceived, modeled using HMMs with parameters trained using isolated audio effects from Web, and decode the optimal sequence using the Viterbi algorithm.

Acoustic information is used also for finding interesting segments of video in video content analysis. Authors of [274] present an audio keyword generation system for sports videos based on audio. They use HMMs for classifying semantic events and a support vector machine (SVM) classifier for finding audio keywords in soccer, basketball and tennis videos. Audio event detection can find a use also in healthcare monitoring for elderly people [194] or audio-based surveillance [56].

The work in [54] deals with direct audio context recognition. Individual events are

considered to be characteristics of the audio scene, and are not modeled themselves, but included in models of the contexts. The events and contexts are chosen such that to minimize overlapping. The authors present results for classifying 14 different contexts using MFCCs and matching pursuit features, using fixed length segments in training and testing.

HMM models are particularly interesting in this context as they consider the transition from one state to another; however it assumes no resampling or permutation of observations.

Although different system architectures and feature sets have been explored [243, 245], even the top rated AED system (around 30% accuracy) left much space for improvement [281]. By contrast, classification of performed isolated events in silent rooms saw very good performance. The evaluation highlighted the challenges in the detection of a large set of ordinary acoustic events in a real world environment [243]. Work reported in [281, 283] tried to further improve the classification performance in a realistic setting with optimisation on extracted features and statistical models. Analysis of the spectral structure of acoustic events and design of a suitable feature set are important for AED. Various audio perceptual features have been proposed for different analysis tasks [39, 43, 221]. In the recent CLEAR Evaluations for AED, the most popular features are speech perception features [14, 243], such as Mel-Frequency Cepstral Coefficients (MFCC) and log frequency filter bank parameters, which have been proven to represent speech spectral structure well.

SVMs were shown to be optimal for classification of isolated events in a silent environments [226], while dynamic Bayesian networks and HMM were applied in noisy environments thanks to the Viterbi algorithm [92], which allows the simultaneous computation for optimal segmentation and classification of the audio stream [242, 243]. Additionally, the use of boosting approaches are recommended to construct a discriminative feature set from a large feature pool.

6.3.3 Feature Extraction for AED

Over the past decades, a lot of research has been done on speech perceptual features [120, 209]. Currently, the speech features are designed mainly based on properties of speech production and perception. Based on knowledge of the human auditory system, the envelope of the spectrogram (formant structure) instead of the fine structure of the

spectrogram (harmonic structure) is believed to hold most information for speech. Both log frequency filter bank parameters and Mel Frequency Cepstral Coefficients (MFCC) [120] use triangular band pass filters to smooth out the fine structure of the spectrogram. Moreover, to simulate the non-uniform frequency resolution observed in human auditory perception, these speech feature sets use bandwidths based on the perceptual critical band, e.g., they have higher resolution in the low frequency part of the spectrum. These features have been successfully used to characterize speech signal as well as other signal perceived by human audition, e.g., music [162].

The spectral structure of acoustic events is different from that of speech, therefore speech feature sets designed according to the spectral structure of speech might be far from optimal for AED [282]; frequency ranges that contain little speech discriminative information, but of great discriminative importance for acoustic events, might be neglected. Even though, AED solutions utilise mostly features extracted from the spectral representation of the signal, some time domain characteristics of the signal such as the zero crossing rate and energy/power, are used, especially for unvoiced and noise like events.

To analyze the spectral structure of acoustic events for AED, Kullback–Leibler Divergence (KLD) based feature discriminative capability analysis may be carried out. KLD allows researchers to understand the relevance of different feature components (in a speech feature set) for the AED task, compared to speech recognition. The distance between the distributions associated with an acoustic event label and the other audio labels reveals the discriminative capability of the feature for that acoustic event.

6.3.4 Acoustic Event Detection for Surveillance

Audio covers a 360° area day and night, at a low cost, and surpasses the limitation of the viewing of conventional surveillance cameras. The lack of audio in surveillance systems impacts significantly the ability of security personnel to act timely, if at all, in cases of emergency. AED aims to identify both time, duration and types of events in an audio stream. This becomes very challenging when going beyond restricted highlight events and well controlled recordings.

As it has been previously discussed in this Chapter, the field of audio signal classification consists of methods for extracting relevant features from a sound in order to identify into which of a set of classes the sound is most likely to fit. The feature extrac-

tion and grouping algorithms used can be quite diverse depending on the classification domain of the application [98].

Research on automatic surveillance systems has recently received particular attention, due to the increasing importance of these systems as well as the prohibitively growing expenses as the number of deployed sensors escalates [256]. In particular, the use of audio sensors in surveillance and monitoring applications has proved to be particularly useful for the detection of events like screaming and gunshots [56] [212]. Such detection systems can be efficiently used to signal to an automated system that an event has occurred and at the same time, to enable further processing like acoustic source localization for steering a video-camera.

Traditional implementations involve the use of speech/music segmentation and classification [164] [197] and audio retrieval [279]. Much of the previous work about audio-based surveillance systems has concentrated on the task of detecting some particular audio events. Early research stems from the field of automatic audio classification and matching [279]. More, recently, specific works covering the detection of particular classes of events for multimedia-based surveillance have been developed. The SOLAR system [122] uses a series of boosted decision trees to classify sound events belonging to a set of predefined classes, such as screams, barks etc.

Successive works have shown that classification performance can be considerably improved if a hierarchical classification scheme composed by different levels of binary classifiers is used in place of a single-level multi-class classifier [14]. The hierarchical approach has been employed in [212] to design a specific system able to detect screams/shouts in public transport environments. A slightly different technique is used in [56] to detect gunshots in public environments. Several binary sub-classifiers for different types of firearms are run in parallel. In this way, the false rejection rate of the system is reduced by a 50% on average with respect to a single gunshot/noise classifier. Finally, in [14] a hierarchical set of cascaded Gaussian Mixture Models (GMM) is used to classify 5 different sound classes. Reported results show that the hierarchical approach yields accuracies from 70 to 80% for each class, while single level approaches reach high accuracies for one class but poor results for the others.

Despite the advances, none of the previously mentioned systems has been developed for operation on computationally and power limiting devices; imposing constraints on the complexity of deployed analytic solutions and the extraction of audio features.

6.4 Validation of Methodology on Audio for Acoustic Event Detection

There is no universal solution for every problem. The design of the methodology to tackle a problem has to serve its specific needs while considering limitations and constraints. The methodology proposed in Chapter 4 has been deployed in [47] an audio classification problem with a twofold objective: first the performance evaluation of a number of algorithms in successfully performing audio event detection and secondly their evaluation in terms of time complexity as the solution aided deployment on a low cost embedded system.

In this task, classifiers of different computational profiles should be produced to allow detection and validation of acoustic events of interest; the desired solution should allow support for two levels of analysis. In the rest of this section, the implementation details of classification solutions for surveillance along with the designed classification strategy will be presented.

6.4.1 Implementation of robust audio analytics for surveillance

Analytics in surveillance systems impose additional restrictions with respect to the complexity of classifiers – to allow their deployment on edge hardware platforms – as well as the classification time. More specifically, the time required for feature extraction, classification and communication of the result needs to be less than the audio block size. Models that do not comply to these restrictions may not be considered as candidate solutions as they may not be operated in real-time.

The methodology presented in Chapter 4 is used for the training of robust classifiers for the detection of gunshots, glass breaking and screaming incidents. The trained models were designed to be deployed as part of an ethical audio surveillance system developed for the cost effective and real-time detection of auditory events of interest. The implemented surveillance system utilises a low-cost embedded system for the recording of audio and first level event detection with the use of lightweight analytics. Extracted features of detected events are transferred to a private cloud for the execution of second level event validation.

As is has already been stated, algorithms are not expected to perform equally well with any dataset. The results are highly dependent on the parameterisation of the

algorithm, as well as the structure and complexity of the dataset. Since models trained for surveillance aim use in open and unrestricted environments, effective operation of the systems – serving the already defined specification – require analysis of the data with multiple techniques where each one examines different aspects of the sample.

6.4.2 System Operational Information

The designed audio analytics solution is composed of two levels of analysis, the models designed for deployment on the edge should allow deployment on resource constraint embedded systems. For the needs of this case study, the following the A10-O LinuXino-LIME equipped with an A10 1GHz Cortex-A8 ARMv7 CPU, 512MB DDR3 RAM memory embedded system was used; a cost effective low-end device which supports wired connections. For the second level of analysis on the cloud side, a dedicated Virtual Machine is allowed a CPU clocked at 2.2 GHz, 2 GB of RAM, and 20 GB of storage capacity. Both embedded and cloud systems run under Ubuntu OS while algorithm implementation is performed in C on the ES and in Python 3.5 on cloud.

The performance of the system is a critical factor, therefore lightweight analytics are performed on the, power limited, embedded system. Upon occurrence of an event the results are transferred on the cloud for further analysis. This method is adopted to limit the number of false positive detection and therefore allowing the system to operate without unwanted traffic.

A model previously trained with the use of sample data is deployed on the power restricted device. Audio is recorded at a sampling rate of 44.1kHz with 16-bit depth, a circular buffer temporarily stores the digitized samples. The necessity to operate and take decisions in real time requires splitting the received data stream into frames of predefined size; each frame is sequentially analysed and a set of extracted features is obtained. Features are extracted from individual audio blocks, to be passed through the previously trained model which is called to make a binary decision indicating or not the existence of an acoustic event (see Figure 6.2). The detection of an event on the first level of detection triggers buffering for successive blocks – even if there are negatively labeled – to allow their aggregation and second level of analysis in the cloud for event validation. Due to the very short length of audio blocks, an event generally generates multiple positively flagged samples. The classifier on the second level of analysis receives batches of samples for prediction.

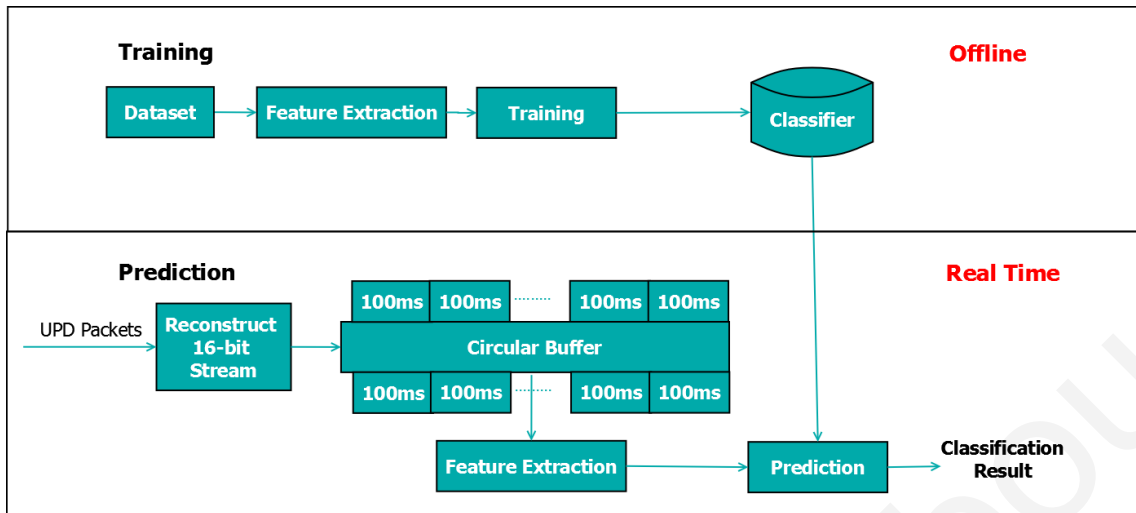


Figure 6.2: Lightweight analytics on ES

The selection of audio analytics models for the first and second levels of analysis followed a comparative study in which re-sampling and boosting, which are integral parts of the methodology, allowed generating new datasets with a balanced number of samples for uniform class distribution. The selection of the preferred classification method in this case study, not only depended on the classification performance, but also on the required processing time. Both parameters are critical as they impact greatly the overall performance of the system. Inadequate classification performance would result in an increase of false positives while slow classification speed would result in overburdening the ES resources resulting in delays and a non-real time solution. During operation, the ES audio analytics module performs feature extraction on discretised blocks of audio of predefined length.

It is expected that an audio event will raise multiple alerts as it will span multiple blocks. Due to this, the second level of analysis implements classification fusion for increased confidence in the process of event detection.

For the needs of this study, a series of audio datasets which do not only involve events of interest but also background noise and other random sounds composing a mixture of different sound sources captured from different environments were used. This process was considered necessary to enhance system robustness and classifiers generalisation capability.

6.4.3 Feature Extraction

Analysis of the audio stream is carried out in a block based fashion where for each block a number of spectral, cepstral and unvoiced coefficients are extracted and used for subsequent classification. For each sound block, the first 13 Mel-frequency cepstral coefficients (MFCC) [202] are retained along with the values of zero-crossings, and block energy.

As cepstral features are computed by taking the Fourier transform of the warped logarithmic spectrum, they contain information about the rate changes in the different spectrum bands. Cepstral features are favorable due to their ability to separate the impact of source and filter in a speech signal. In other words, in the cepstral domain, the influence of the vocal cords (source) and the vocal tract (filter) in a signal can be separated since the low-frequency excitation and the formant filtering of the vocal tract are located in different regions in the cepstral domain. If a cepstral coefficient has a positive value, it represents a sonorant sound since the majority of the spectral energy in sonorant sounds are concentrated in the low-frequency regions. On the other hand, if a cepstral coefficient has a negative value, it represents a fricative sound since most of the spectral energies in fricative sounds are concentrated at high frequencies. The lower order coefficients contain most of the information about the overall spectral shape of the source-filter transfer function. In particular, the zero-order coefficient indicates the average power of the input signal and the first-order coefficient represents the distribution spectral energy between low and high frequencies. Even though higher order coefficients represent increasing levels of spectral details, depending on the sampling rate and estimation method, 12 to 20 cepstral coefficients are typically optimal for speech analysis. Selecting a large number of cepstral coefficients results in more complexity in the models.

6.4.4 Classifier Training Methodology

The designation of the implemented methodology accounts for the non-uniformity in the distribution of classes, as well as the fact that audio samples significantly different from the training set may be collected during a real-life scenario. For the needs of this study, bootstrapping with replacement is used to allow the generation of new datasets; a balance between true and false events was maintained to avoid over fitting to either category. The choice of re-sampling adds the assumption that the dataset

is representative to the population. Moreover, the performance of the algorithms is evaluated using the classification accuracy, i.e. the number of correct predictions from all predictions [232], as well as the AUC index [38]. The calculation of both indices is done in an effort to assess their robustness and emphasise the significance of their choice.

The 5x2 cross validation paired t-test [68] and the 5x2 cross validation F-test [10] were deployed to statistically test the significance of the classification results and to evaluate their robustness. Benchmarks on significance testing propose that cross validation testing methods are more robust when dealing with small datasets where reproducibility of the experiment is not an issue [68]. The 5x2 cross validation method was selected to allow large enough datasets for testing while ensuring that no further dependencies of overlapping training and testing sets are introduced when cross validation is used [217].

Variants of the scheme introduced in Chapter 4 has been deployed for the training of classifiers for both analysis levels to serve different objectives. On the cloud, the objective was to determine which algorithm performs the best, in terms of classification evaluation metrics, while on the embedded system the objective was to find the algorithm that satisfies the trade-off between classification performance and real-time operation. In each instance, the experiment was allowed to run for 500 iterations to allow the generation of valid statistics and the significance of the results was calculated at level 0.05 ensuring that a 95% confidence level for the results of statistical testing.

6.4.5 Classification Algorithms used in the experiment

The selection of the algorithm was determined based on the results of an extensive experiment involving a number of classification methods some of which were also used in Chapter 5.

Random Forests

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [35] [259]. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large while it also depends on the strength of the individual trees in the forest and the correlation between them.

Random forests on bagging where the combination of learning models, is thought to, increase the classification accuracy. The main idea of bagging is to average noisy unbiased models to create a model with low variance, in terms of classification; a large collection of decorrelated decision trees. Supposing that matrix S is the set of training samples, used for training the classification model, random forests generate a number of different and independent decision trees from an equal number of random data subsets (selected as random sets from the original set. Upon prediction, new uncategorised samples are classified by all trained trees. The element is assigned the label based on the majority rule of class label each tree has returned.

Support Vector Machines

Support vector machine is based on statistical learning theory and the structural risk minimization principle [259]. Using the training data, SVM implicitly maps the original inputs space into a high dimensional feature space [140]. Subsequently, in the feature space the optimal hyperplane is determined by maximizing the margins of class boundaries [1].

The training points that are closest to the optimal hyperplane are called support vectors. Once the decision surface is obtained, it can be used for classifying new data. Consider a training dataset of instance-label pairs (x_i, y_i) with $x_i \in \mathbb{R}^n$, $y_i \in \{1, -1\}$ and $i = 1, \dots, m$. In the current context of audio classification, x is a vector of input space that contains the previously extracted audio coefficients. The two classes 1,-1 denote identified event and misidentified event. The aim of the SVM classification is to find an optimal separating hyperplane that can distinguish the two classes $\{1, -1\}$ from the mentioned set of training data. For the case of linear separable data, a separating hyperplane can be defined as: $y_i(w \times x_i + b) \geq 1 - \xi_i$ where w is a coefficient vector that determines the orientation of the hyperplane in the feature space, b is the offset of the hyperplane from the origin, ξ_i is the positive slack variable [60].

k-Nearest Neighbour (k-NN) & C4.5

The k-NN and C4.5 classification algorithms were used in Case Studies. These methods were briefly introduced in Section 5.4.3. The k-Nearest Neighbour (k-NN) implementation in this case study was used with Manhattan distance.

6.4.6 Lightweight Audio Analytics

The audio analytics module on the embedded system (ES) is designed to emphasise on the detection of screaming, glass breaking and gun-shooting/loud explosions with the deployment of a timely efficient method.

Within the tasks of the embedded system is the recording of audio from the attached microphone, its encoding in the WAVE uncompressed format, its re-sampling to the preferred sampling rate as well as the transmission of extracted features and lightweight analytics results on the cloud, therefore the latency caused by these aspects needs also to be considered.

Additionally, due to privacy reasons, the audio data should neither be retained on the ES after a block has been processed, nor be transmitted on the cloud through the network. Therefore, only the extracted feature coefficients are transmitted to the cloud for further processing (see Figure 6.3). The need of operating in real-time imposes the extra challenge in preserving low processing (feature extraction and classification) times.

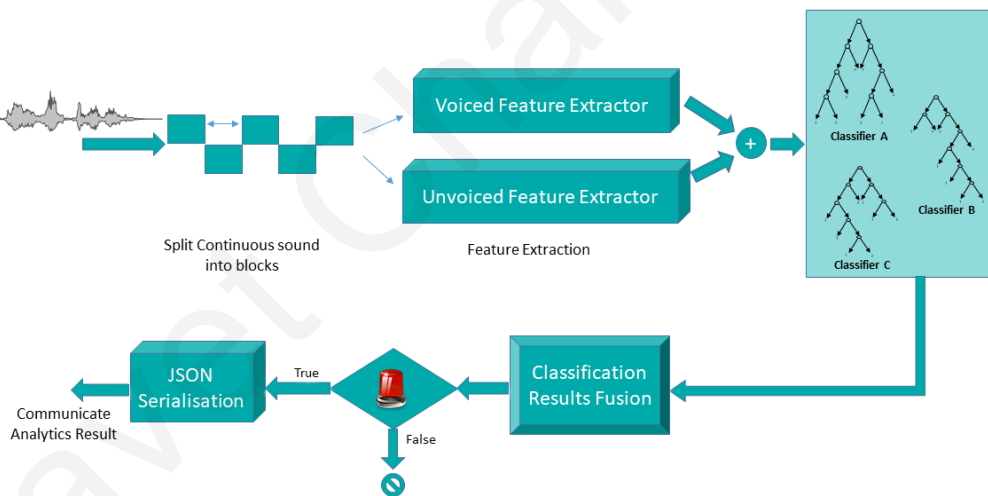


Figure 6.3: Embedded System Audio Analytics Functional View

Classification on the ES is performed with C4.5. The selection of this method emerged after a systematic experiment discussed in [46]; during this process the effectiveness of a number of possible audio features has also been tested. Statistical testing revealed that in most cases C4.5 was significantly better than its counterpart algorithms while in cases where the null hypothesis (i.e. algorithm A is not significantly

better than algorithm B) is confirmed C4.5 was proven to be more efficient in terms of processing times.

Time Complexity

The necessity to operate and take decisions in real time requires splitting the received data stream into frames of predefined size; each frame is sequentially analysed and a set of extracted features is obtained. The prediction module is then called to make the binary decision (i.e. the sound is alarming or not). The trade-off between processing time and classification error (measured as the misclassification rate) is considered critical and as a result, the parametrisation of the algorithm is determined based on the results, shown in Table 6.1, involving the following parameters: audio sampling frequency, block size (expressed in ms), NFFT frame, number of filter bank and MFCC coefficients, misclassification error and average (block) processing time. Based on the reported times in Table 1 the parameterisation of configuration #1 is adopted.

Table 6.1: Top 3 configurations obtained in an experiment which involved the parameterisation of the audio analytics module.

Configuration			
Parameter	#1	#2	#3
$F_s(\text{kHz})$	8	8	8
$Block(\text{ms})$	140	100	140
$FBank$	22	22	22
$MFCC$	13	13	10
$Accuracy$	98.6%	98.4%	98.4%
$Time(\text{ms})$	85.74	89.06	65.13

Classification strategy

The audio analytics module on the embedded system solves the classification problem with a number of cascaded binary classification trees (shown in Figure 6.4); generated with the means of the C4.5 algorithm. Processing involves a number of steps depending on the output of previous steps. First the input sample goes through the three trees in Tier 1, in the case where all algorithms classify the sample as non-alerting no further processing is required, otherwise processing goes through the tree in Tier 2 which

identifies between voiced and unvoiced sounds. If a scream is detected then the sample is labelled as screaming and the results along with features is queue for transmission to the cloud. In the event of an unvoiced event the algorithm moves to Tier 3 and the discrimination between the sounds of glass breaking and gunshots.

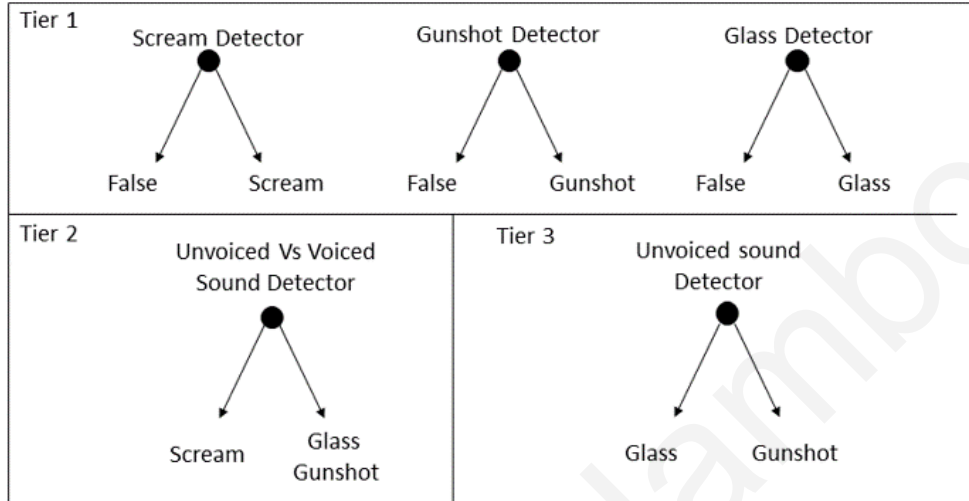


Figure 6.4: Lightweight analytics on ES

6.4.7 Cloud side analysis

On the cloud side a number of different analytics methods are available for analysis. Due to the diversity of audio sounds that may occur in trial setups, the analysis of the data from different perspectives is considered necessary, therefore the selection of the algorithms was driven by this need; each algorithm exploits best certain data attributes. Despite our selection of classification methods, the proposed design supports the addition of any other classification/ clustering method.

Due to this, four different methods have been deployed: C4.5 based on decision trees, random forests, an ensemble DT method, and two statistical methods known as Support Vector Machines (hereafter SVM) and k- Nearest Neighbour (hereafter k-NN). A more complex structure, based on neural networks, known as Learning Vector Quantisation (LVQ) [261] has also been implemented, however during testing it was decided to be left out as the produced results were inconsistent.

Due to the continuous nature of sound waves and the fact that an acoustic event is very likely to span multiple processing blocks, a varied majority voting approach is also applied on the classifier's predictions prior to the final output. This step has been added to the event validation process as a countermeasure against spontaneous false

positives. Majority voting is applied as a means of fusing classification results when multiple classifiers produce predictions on the same data.

Two different voting schemes are common among voting classifiers, hard and soft voting. In hard voting, every individual classifier votes for a class, and the majority wins. In statistical terms, the predicted target label of the ensemble is the mode of the distribution of individually predicted labels. A hard majority vote classifier consisting of votes from hypotheses h_1, h_2, \dots, h_B is defined as follows.

$$C(X) = \arg \max_i \sum_{j=0}^B w_j I(h_j(X) = i) \quad (6.1)$$

where w_1, \dots, w_B are weights that sum to 1 and $I()$ is an indicator function. Data samples x are assigned to the class that receives the largest number of classification votes. In soft voting, every individual classifier provides a probability value that a specific data point belongs to a particular target class. The predictions are weighted by the classifier's importance and summed up. Then the target label with the greatest sum of weighted probabilities wins the vote.

$$C(X) = \arg \max_i \sum_{j=0}^B w_j \hat{p}_{ij} \quad (6.2)$$

where \hat{p}_{ij} is the probability estimate from the j^{th} classification rule for the i^{th} class.

In the context of the task in hand, the transferred data from lightweight analytics that represented an acoustic event was provided as input to a trained classifier. The estimated classification output was applied to soft majority voting with equal weights to produce the validated classification outcome.

Classifier Performance Evaluation

The implemented cloud analytics algorithms for audio offer more flexibility compared to the designed strategy on the ES for lightweight audio analytics. The computational resources on the cloud enable the quick analysis of transferred parameters through multiple algorithms in low confidence cases. The same approach as earlier has been followed for the evaluation of algorithms on cloud. During the experiment all algorithms operated with the same parameterisation for all generated datasets so as to avoid the introduction of dataset bias in the performance of the algorithms, however these numbers emerged as a result of several experiments. Therefore, C4.5 operated with no pruning, k-NN with k=3, random forests with tree bagging of 20 and finally

SVM with the RBF kernel, a maximum value of 15000 iterations and a 5% level of violation of the KKT conditions in cases where the algorithm does not reach convergence. For consistency purposes the evaluation of each algorithm for each type of alert involved the calculation of the following metrics: Accuracy, Sensitivity, Specificity, Precision, Recall, F-Measure, G-Mean, and AUC. In total a number of three experiments have been performed, one for each of the three possible alerts; results shown in Table 6.2. The experiments, tested the algorithms performance in discriminating the sounds of glass breaking, gunshots/explosions and screaming from non-scream sound (background noise, people talking, ambient sound in transportation media etc.).

Table 6.2: Classification performance for the detection of events.

	Glass Vs. NonScreams				Gunshot Vs. NonScreams				Screams Vs. NonScreams			
	C4.5	k-NN	RF	SVM	C4.5	k-NN	RF	SVM	C4.5	k-NN	RF	SVM
Accuracy	1.00	1.00	1.00	0.81	0.99	0.98	0.99	0.97	1.00	1.00	1.00	0.99
Sensitivity	1.00	1.00	1.00	0.78	0.99	0.99	0.99	0.97	1.00	1.00	1.00	0.99
Specificity	0.99	1.00	1.00	1.00	0.86	0.93	0.96	0.99	0.97	0.98	0.99	1.00
Precision	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	0.78	0.99	0.99	0.99	0.97	1.00	1.00	1.00	0.99
F-Measure	1.00	1.00	1.00	0.87	0.99	0.99	1.00	0.99	1.00	1.00	1.00	0.99
Gmean	0.99	1.00	1.00	0.88	0.93	0.96	0.97	0.98	0.98	0.99	0.99	0.99
AUC	1.00	1.00	1.00	0.71	0.93	0.86	0.94	0.75	0.98	0.92	0.99	0.70

6.5 Experiment Results & Conclusions

Results of the above experiment suggested that C4.5 was consistently the quickest algorithm both during prediction and training while it also returned high classification scores. k-NN performs best in discriminating between the sound of glass breaking and non-screaming, even though RF and C4.5 follow its performance very closely; k-NN is no significantly better than C4.5 and RF with all 3 algorithms reporting extremely high rate in both accuracy, F-measure and AUC. RF outperforms the rest algorithms in the detection of gunshots and abnormally loud sounds reporting almost perfect classification, very shortly followed by C4.5 and k-NN. However the results show that RF and C4.5 report better AUC values than k-NN. RF outperforms also in the detection of screams with very high rates, shortly followed by C4.5 and k-NN. Despite this, none of the algorithms is significantly better than the others in terms of Accuracy, Sensitivity, Specificity, Precision, Recall, F-Measure and G-Mean. SVM constantly produces lower performance rates compared to the rest algorithms in terms of AUC. The variability

between reported values of accuracy, f-measure and AUC provide a lead for further investigation of characteristics between the different classes. SVM was parametrised in a way to accept a degree of error in favour of running time. The parameterisation allows for extremely quick classification.

The fully automatic and reliable identification of sounds and alerts in real-time, at the current stage of advancement in technologies, is not possible when a computationally restricted ES is involved in the process. However, the results indicate that reliable ways of detecting alerts within an environment are feasible. Despite the high scores of the performed experiment, a confidence classification metric may be calculated as a function of the number of alerts that have occurred in the clip, the matching between ES and cloud analytics as well as the fusion of results obtained from the simultaneous analysis of the clip with multiple algorithms (possibly audio, video and/or depth).

Chapter 7

Data Analysis Suite Tool

7.1 Introduction

Data analytics tools can help deliver value and bring data to life. A lot of hard work goes into extracting and transforming data into a usable format, but once that's done, data analytics can provide users with greater insights into their customers, business, and industry. Tens of tools, some of which open sourced or freely available, have been launched over the past decade to assist data analysis experts and non to gain insights in their data with the aim to either extract valuable knowledge or to introduce automations in existing or new systems. As data analysis comes in many forms and covers a wide spectrum of operations, applications such as: Tableau Public, Rapid-Miner, KNIME, QlikView and Splunk serve as mere examples on the range of available tools that offer such services for specific data types and sources. On the other hand, technologies such as: Apache Spark, Scikit Learn and TensorFlow enabled by R, Matlab and Python environments allow the generation and configuration of advanced and parametrised data analysis models to fit the requirements of any problem.

However, research presented in Chapter 3 illustrates that many parameters need to be taken into consideration during the preprocessing, training, validation, deployment and maintenance phases of data analysis models for robustness in the produced outcome.

7.2 Motivation

The use of ML in so many domains and applications has led to an emerging need of application domain experts to, not only familiarise themselves with the basic principles of data analysis, but also to gain substantial comprehension on data sampling, transformation and analysis practices. Considering, the complexity of this task it is rather challenging to train robust classification models with validated performance without coding or scripting involved.

The aforementioned challenge impacts the sectors of research and industry and escapes beyond the areas of archaeology and security which we analysed as case studies. As a result, a web-based data analysis suite tool have been developed to allow application domain experts to gain insights on their data, to train classifiers with assured performance and to deploy with unseen data. The rest of this chapter discusses the features and functionalities covered by this tool.

7.3 Tool Design

The data analysis suite tool has been developed in an effort to allow users to gain insight on their data quickly through an easy to use interface. The data analysis platform covers the following functionalities to solve binary problems:

- Data importing either in raw media format for audio data or in CSV format for any other type of numeric dataset,
- Exploratory data analysis with clustering methods on previously imported data,
- Model training and validation for classification problems,
- Classification of unseen data with previously trained models.

The data analysis suite is being developed based on open-source technologies and the Django web-development framework [123] on an MVC design (see Figure 7.1), with Python powering the server-side and HTML, CSS, Javascript and JQuery on the client-side. The user interface is accessible through any modern web browser and does not require any scripting knowledge. For ease of deployment, the project is packaged in two docker containers: one for the web service and one for the tool's database.

For flexibility and scalability purposes, the processing and visualisation layers are designed as separated layers that communicate through a REST 2.0 API. Configuration

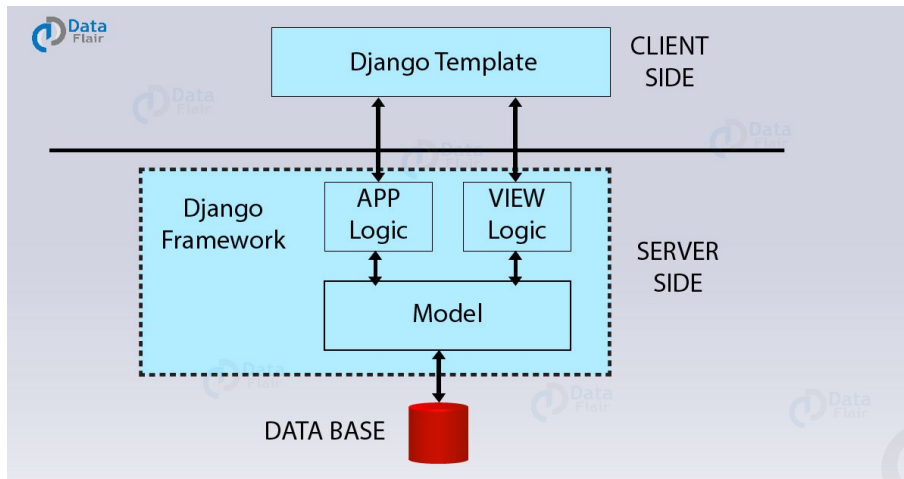


Figure 7.1: Django Architecture diagram. [104]

and admin data are stored in a secured Sqlite3 database while extracted features and analysis results are stored in a Mongo database to allow storage and management of large volumes of data(see Figure 7.2).

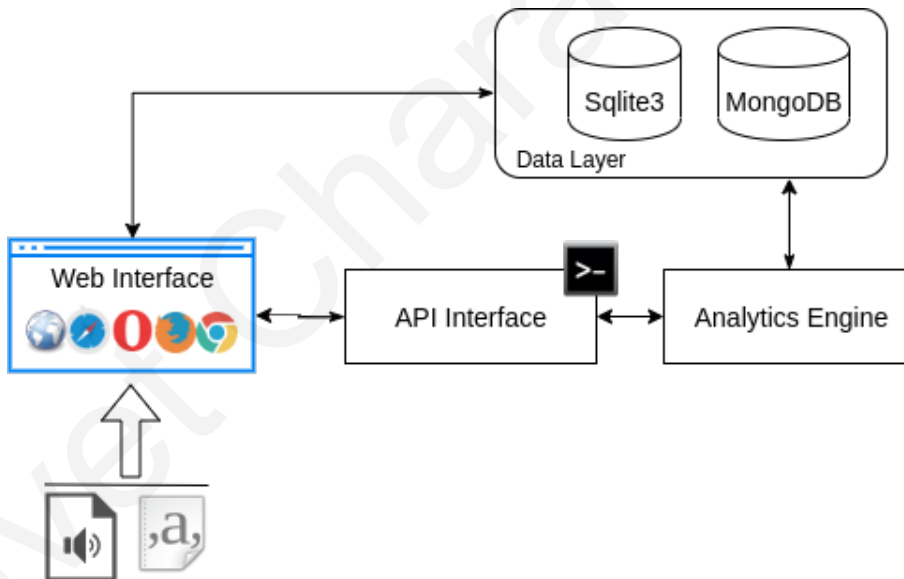


Figure 7.2: Conceptual Design

The data analysis platform has been implemented with standardised interfaces to allow a software developer to easily add new classification and clustering algorithms as well as to add upon the already provided parametrisation options. At the current stage to allow for testing, the data analysis platform integrates the SVM [60], Random Forests [35], Multilayer Perceptron Neural Networks [96], Naive Bayes [277] and Stochastic Gradient Descent [32] classification algorithms and the K-Means [130] and Affinity Propagation [94] algorithms for clustering.

7.4 Analysis Principles

The analysis domain of the developed tool relies on the fact that the user will import representative data to allow for adequate class separation upon training, validation and testing . The administrator's panel allows the configuration of class labels as well as validation/evaluation and algorithmic method characteristics as any time.

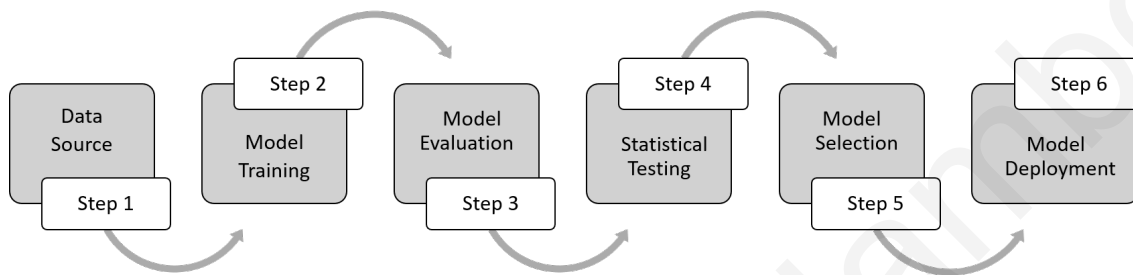


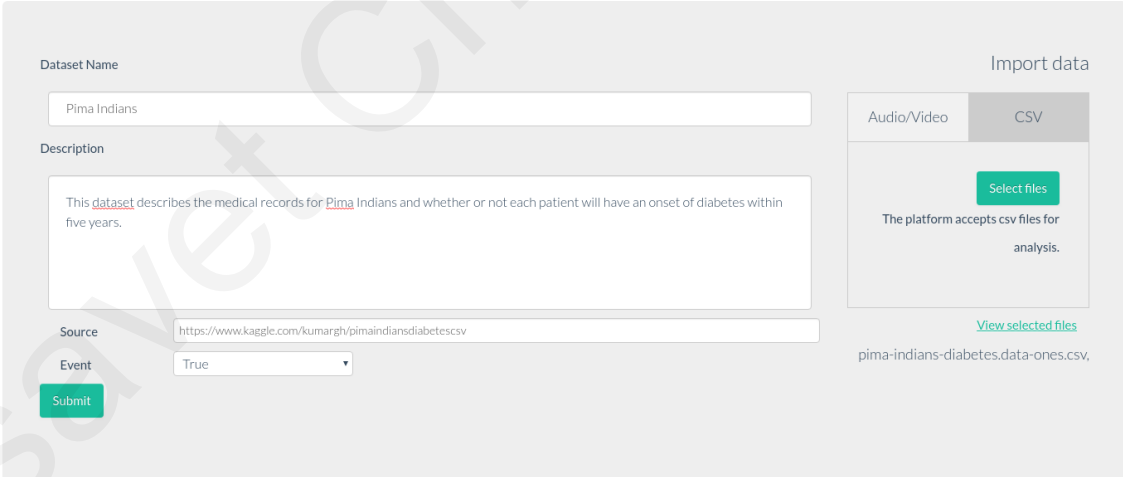
Figure 7.3: Analysis Domain

One of the main requirements was to provide a solution that is dataset and classification method agnostic, the user is called to import new data and associate these with already configured class label . Depending on the type of data, the platform performs pre-processing for data harmonisation purposes and extracts feature domination characteristics . Imported data may become subject to clustering and dimensionality reduction for exploratory data analysis purposes or to model training for classification purposes. The latter option, which is also the main purpose of the tool, allows automated training of multiple models through different and diverse classification techniques. Prior to clustering and classification operations the user is able to select whether feature selection techniques, based on feature ranking, will be applied to perform analysis with the use of selected features. Model performance evaluation is performed based on a series of metrics, while statistical testing may be applied to compare classifier performance for configurable significance indices . Finally, adequately performing models may be selected for use with new data or exported for use independently or in combination with other tools; the latter option is provisioned for cases where a user may perform cascading on trained classification models. This process is summarised in Figure 7.3.

7.4.1 Data Source Processing

Users are able to add import datasets through the Data Input Panel either by uploading CSV files or by uploading a series of audio or video files supported by FFMPEG¹. The data analysis platform is implemented on the concept that a sample feature vector is consisted of numeric values. In CSV files, each data row represents a sample while the first row is thought to be the column headers. In the case where an acoustic dataset is to be imported, the platform transcodes all files in the lossless WAVE format, in mono sound with 16kHz sampling rate; these values were selected as a result of experiments implemented within the context of Chapter 6. Each audio file is processed in 1 sec blocks (with this value being configurable). From each block a range of 193 features are extracted including 40 MFCCs, 12 mean Chromagram values [83], 128 mean mel-scaled spectrogram values, 7 spectral contrast values [134] and 6 tonal centroid features (tonnetz) [112]. The Chromagram and spectral contrast for each block are estimated based on the Short-Time Fourier Transform (STFT) representation of the signal, a time-frequency domain representation computed as a result of discrete Fourier transforms (DFT) over short overlapping windows.

Data Input Panel



The screenshot shows a web interface for creating a dataset. On the left, there is a form with the following fields: 'Dataset Name' (text input with 'Pima Indians'), 'Description' (text area with a placeholder text), 'Source' (text input with a URL), and 'Event' (dropdown menu with 'True'). A 'Submit' button is at the bottom left. On the right, there is an 'Import data' section with two tabs: 'Audio/Video' and 'CSV'. The 'CSV' tab is active, showing a 'Select files' button, a note 'The platform accepts csv files for analysis.', and a 'View selected files' link. Below this, the filename 'pima-indians-diabetes.data-ones.csv' is displayed.

Figure 7.4: Data Input Panel

Other data recorded along with the extracted features during the importing of a dataset involve the dataset name, description, source and associated class label (see Figure 7.4). Once source files are processed, the uploaded files are removed from the server-side and only the extracted features are retained; this was implemented to ensure

¹<https://www.ffmpeg.org/>

that no personal data are retained and no reconstruction of the original data is possible when the tool was applied to surveillance data.

Dataset Details

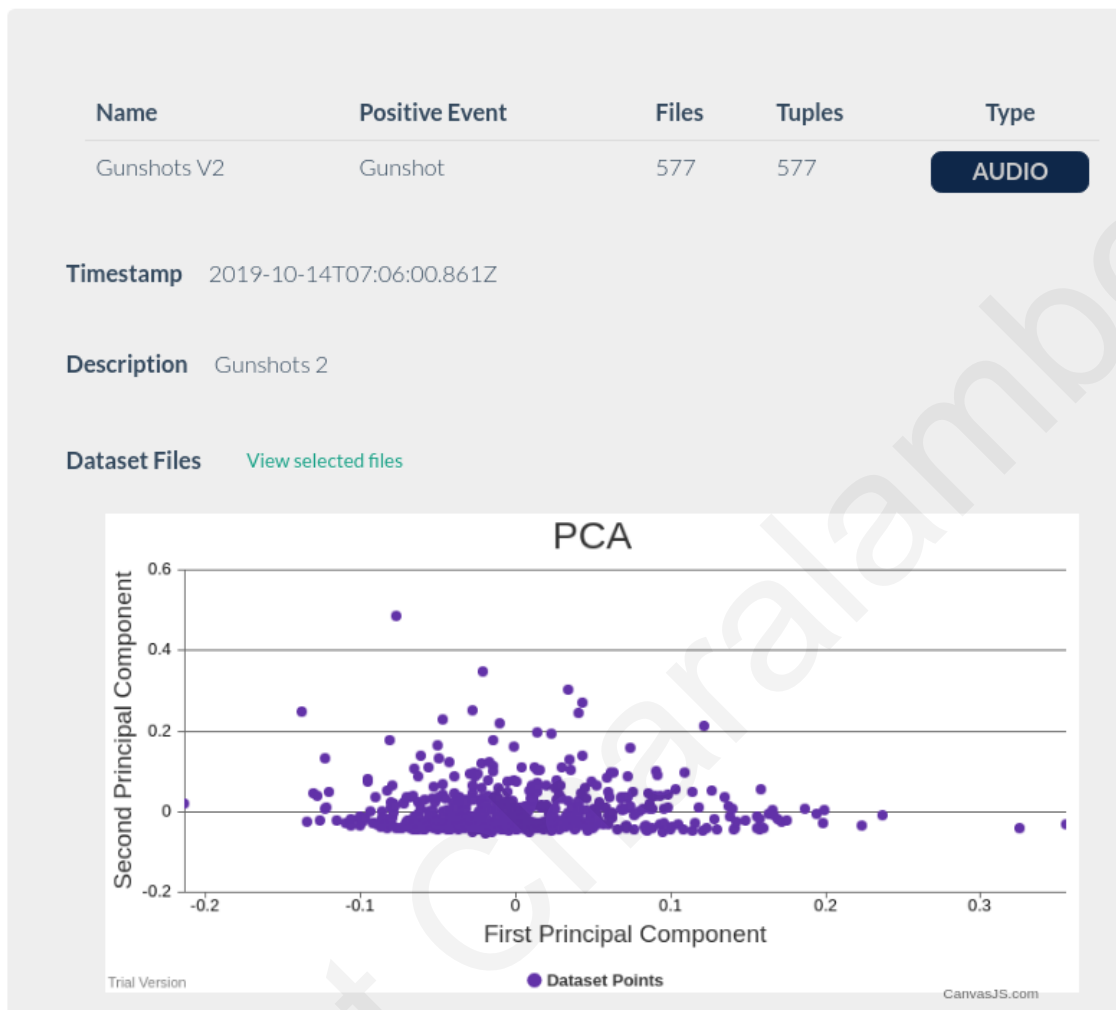


Figure 7.5: Imported dataset detail view

In addition, to the extraction/parsing of dataset features, the Data Analysis platform performs dimensionality reduction based on Principal Component Analysis (PCA) [269] for visualisation purposes. Additionally, the explained variance of the first 10 components (if available) as well as feature domination and normalised feature domination are visualised to allow the user to gain insight on the imported dataset without altering its characteristics. A sample view of an imported dataset is shown in Figure 7.5.

7.4.2 Exploratory Data Analysis

Exploratory data analysis functionalities through dimensionality reduction and clustering methods have been added to the data analysis platform to allow users to review

characteristics relevant to the imported datasets. This is particularly important when data are scarce and analysts are obliged to use multiple datasets collected separately with different recorded methods in order to produce a representative set of samples.

The exploratory data analysis panel allows the user to select one or more datasets to perform clustering or to review the results of previous clustering requests. The user interface allows for parametrisation of the already integrated clustering methods with dynamically updated selections while it also supports for feature selection with SelectKBest and the χ^2 score function independent of the predictive method.

Functionality provided through the EDA panel allow users to review various characteristics of their data along with the impact of feature selection in increasing separability between clusters (see Figure 7.6). Currently, the K-means and Affinity propagation clustering algorithms are supported within the data analysis platform; along with several parametrisation options through the user interface. The selection of the integrated clustering methods was so to allow also for the prediction of the number of clusters. Further clustering methods may be added programmatically through standardised interfaces implemented in the analytics engine component (see Figure 7.2).

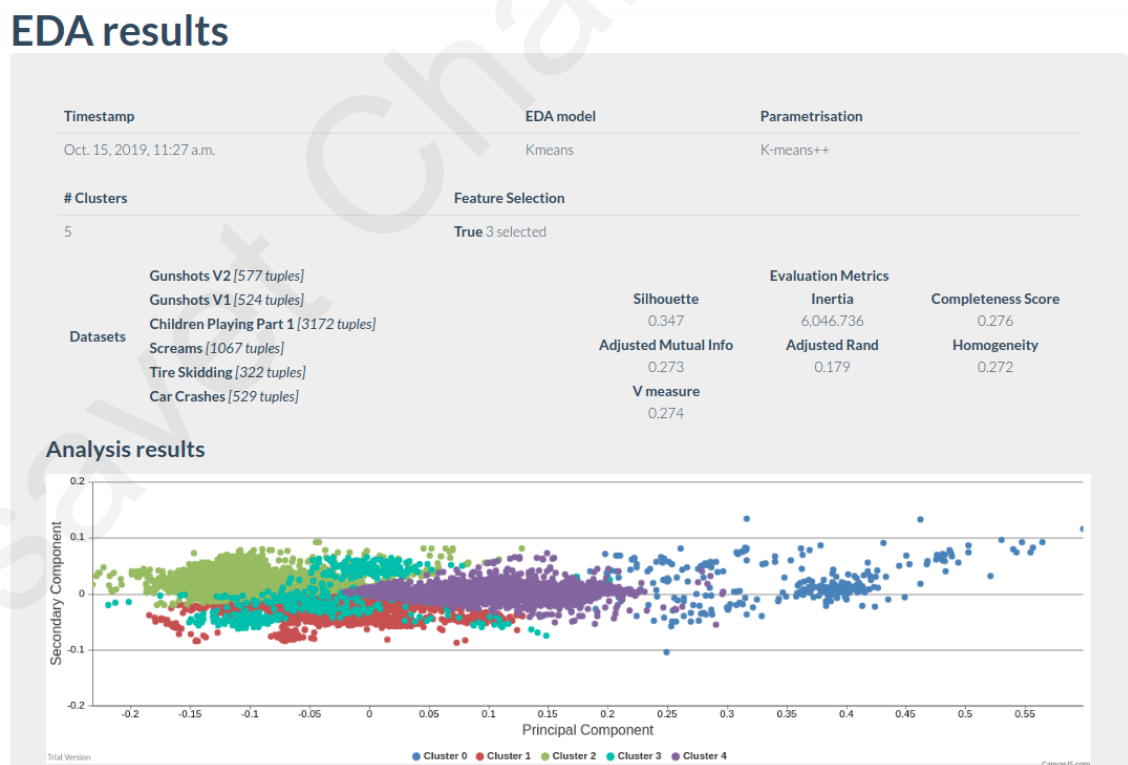


Figure 7.6: Detailed view of clustering result

Once the clustering process is completed, a new entry is added in the list of results. The detailed view of each clustering request shows a PCA representation of

the clustered data labeled with their assigned cluster label. Additionally, a number of evaluation metrics measuring cluster compactness and separation are reported in addition to plots for the explained variance of the first ten features, feature domination and normalised feature domination.

7.4.3 Model Training & Analysis for Classification Problems

The initial driver for the ideation of the data analysis platform was the need for domain experts to be able to train robust classifiers with their own data for later use with unseen data in a user friendly environment. It was also observed that it is difficult to compare classifier performance with when either the training or testing data are different. Due to this, the data analysis suite tool implements two different spaces: the model training panel and the post analysis panel.

The model training panel is designed to allow the user to train multiple classifiers, based upon selection of integrated algorithms, and to evaluate their performance according to the methodology introduced in Chapter 4. As a consequence, the post analysis panel is implemented to allow the classification of unseen data with the use of previously trained and saved classifiers.

Model Training

The model training panel allows the user to submit requests for training multiple classifiers in a single request and to evaluate their performance on the basis of the same data (see Figure 7.7). In particular, the user is called to select datasets for positive and negative detection (binary classification problems) and to pick from a range of already integrated classification algorithms; these include SVM, Random Forests, Multi-layer Perceptron Neural Network, Naive Bayes and Stochastic Gradient Descent. Within the mandatory parameters is also the evaluation metric². The Data Analysis Suite then allows for several cross validation options, feature selection with random forests as well as significance testing with the statistical T-test and F-Test methods. Additionally, in order to avoid cases of dataset dominance due to the vast difference in the size of positive labelled and negative labeled samples, the option of random data truncation is implemented to allow this on an ad-hoc basis even after dataset importing.

²One of ROC AUC, Accuracy, Cohen's Kappa, Confusion Matrix, Jaccard Similarity, Precision, Recall and ROC Curve.

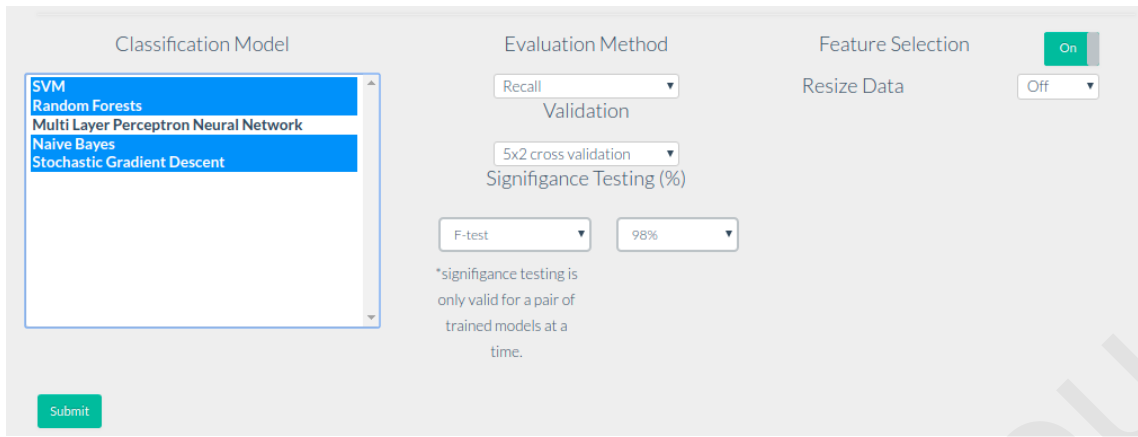


Figure 7.7: Snippet of training panel parametrisation

The submission of a training request entails a number of steps to be performed by the analysis engine component of the tool summarised in Figure 7.8. Firstly, the features from the already processed datasets are retrieved from MongoDB and aggregated to form the positively and negatively labeled datasets when multiple datasets have been selected. At this stage, if data resizing is selected the analysis engine randomly selects x features from the larger dataset where $x = \min(a, b)$ and a and b are the number of samples in the positively labeled and negatively labeled datasets; model training is performed on the selected x samples than on the original dataset.

If feature selection is selected, the new aggregated dataset which combines samples from both class labels, undergoes feature ranking and selection with the ensemble method of Random Forests. Random Forests are often used for feature selection in a data science workflow. The reason is because the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node. This mean decrease in impurity over all trees (called gini impurity). Nodes with the greatest decrease in impurity happen at the start of the trees, while notes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, a subset of the most important features is retained. Following empirical testing with respect to generalisation and ability to cope with heterogeneous datasets, features with importance greater than 1% are retained.

Feature selection is an effective method to increase separability between classes especially when samples are characterised by hundreds or even thousands of features. Generating, simple to interpret, models that only consider important features also reduces the model's variance and computational cost of training.

Following feature selection the training model training process proceeds according

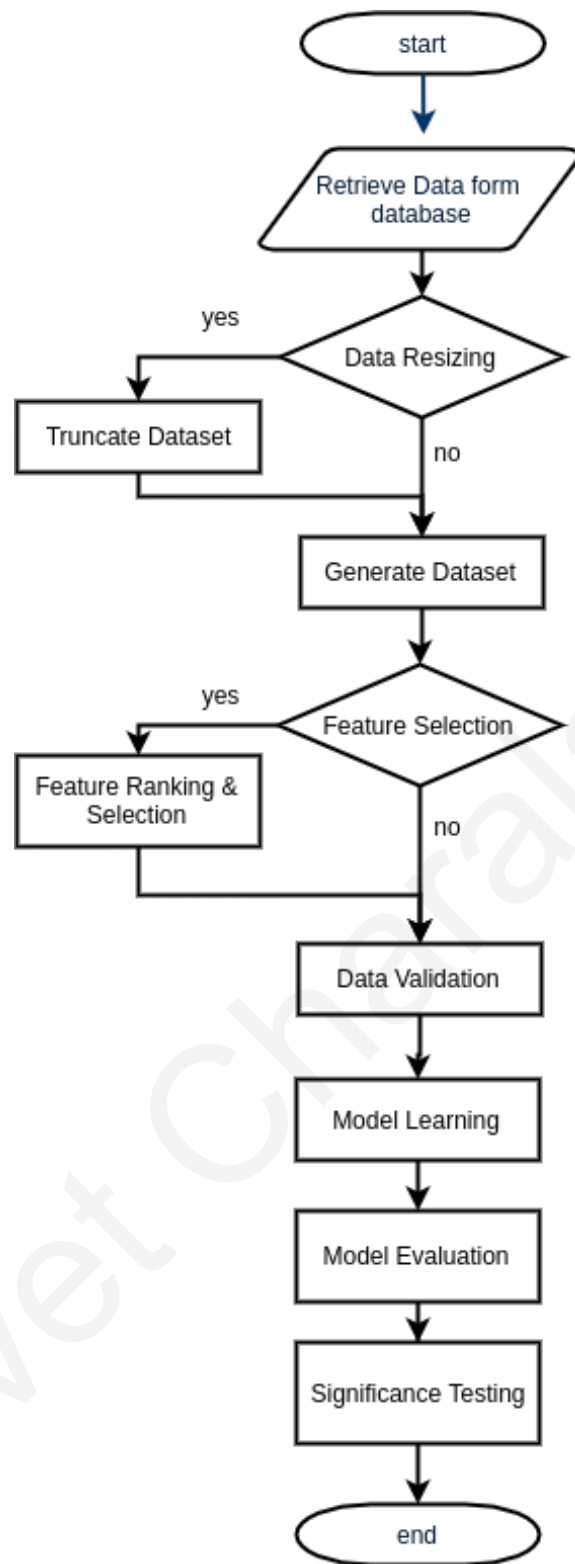


Figure 7.8: Model training through the data analysis suite.

to the data validation parameter selected by the user. The following validation options are available through the tool:

- Data split options for 60-40 data-split, 70-30 data-split and 50-50 data-split where data corresponding to the first percentage are used for training and the percentage

for the second percentage are retained for testing and validating and performance of the trained model;

- $S - fold$ cross-validation: with options for 2-fold and 10-fold cross-validation. $S - fold$ involves taking the available data and partitioning it into S groups (in the simplest case of the same size). Then $S - 1$ of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is then repeated for all S possible choices for the held-out group. The performance scores from the S runs are then averaged [76];
- Leave One Out: This is an extreme case of cross-validation when data is particularly scarce, it may be appropriate to consider the case where $S = N$, where N is the total number of data points [76];
- 5x2 cross-validation: In this test, 5 replications of 2-fold cross-validation. In each replication, the available data is randomly partitioned into two equal-sized sets S_1 and S_2 . Each learning algorithm is trained on each set and tested on the other set. [66];

Depending on the classification problem, the classification performance may be measured through a number of metrics. The data analysis tool allows measuring a classifier's performance with the calculation of one of the following of which one is to be selected prior to the submission of a training request:

- ROC Curve: A receiver operating characteristic curve (ROC curve) is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection [79];
- ROC AUC: An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC [89]
- Accuracy: In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated [124];

- Cohen's Kappa: The Kappa statistic is used to measure the agreement between two sets of categorizations of a dataset while correcting for chance agreements between the categories. The Kappa statistic makes use of both the overall accuracy of the model and the accuracy within each category, both in terms of the predictive model and the field-surveyed sample points, to correct for chance agreement between categories [133].
- Confusion Matrix: The basic confusion matrix is a $k \times k$ matrix of counts, where k is the number of classes involved in the classification problem. By weak tradition, the columns correspond to the true classificatory state, while the rows correspond to the algorithm results. If an object is truly of class j and the algorithm classifies it into class i , then the count in cell (i, j) —the cell at the intersection of row i and column j —of the confusion matrix is incremented by one [91].
- Jaccard Similarity is included in the negative match exclusive measures and is calculated on the basis of the Operational Taxonomic Units (OTUs) [86] in a 2×2 contingency table. The Jaccard coefficient proposed at 1901 is still widely used in the various fields such as ecology and biology [53].
- Precision: used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class [124];
- Recall: Recall is used to measure the fraction of positive patterns that are correctly classified [124];

Finally, the 5×2 cross validation paired t-test [67] and the 5×2 cross validation F-test [10] can be deployed, if selected, to statistically test the significance of the classification results and to evaluate their robustness. Benchmarks on significance testing propose that cross validation testing methods are more robust when dealing with small datasets where reproducibility of the experiment is not an issue [68]. The 5×2 cross validation method was selected to allow large enough datasets for testing while ensuring that no further dependencies of overlapping training and testing sets are introduced when cross validation is used [217].

This functionality has been added to alleviate the factor of likelihood in an algorithm's performance against another algorithm and can be deployed only for pairs of classification algorithms. The tests essentially check the probability that the selected

evaluation statistic has a high enough probability of being drawn from that distribution [10].

Overview of Trained Models

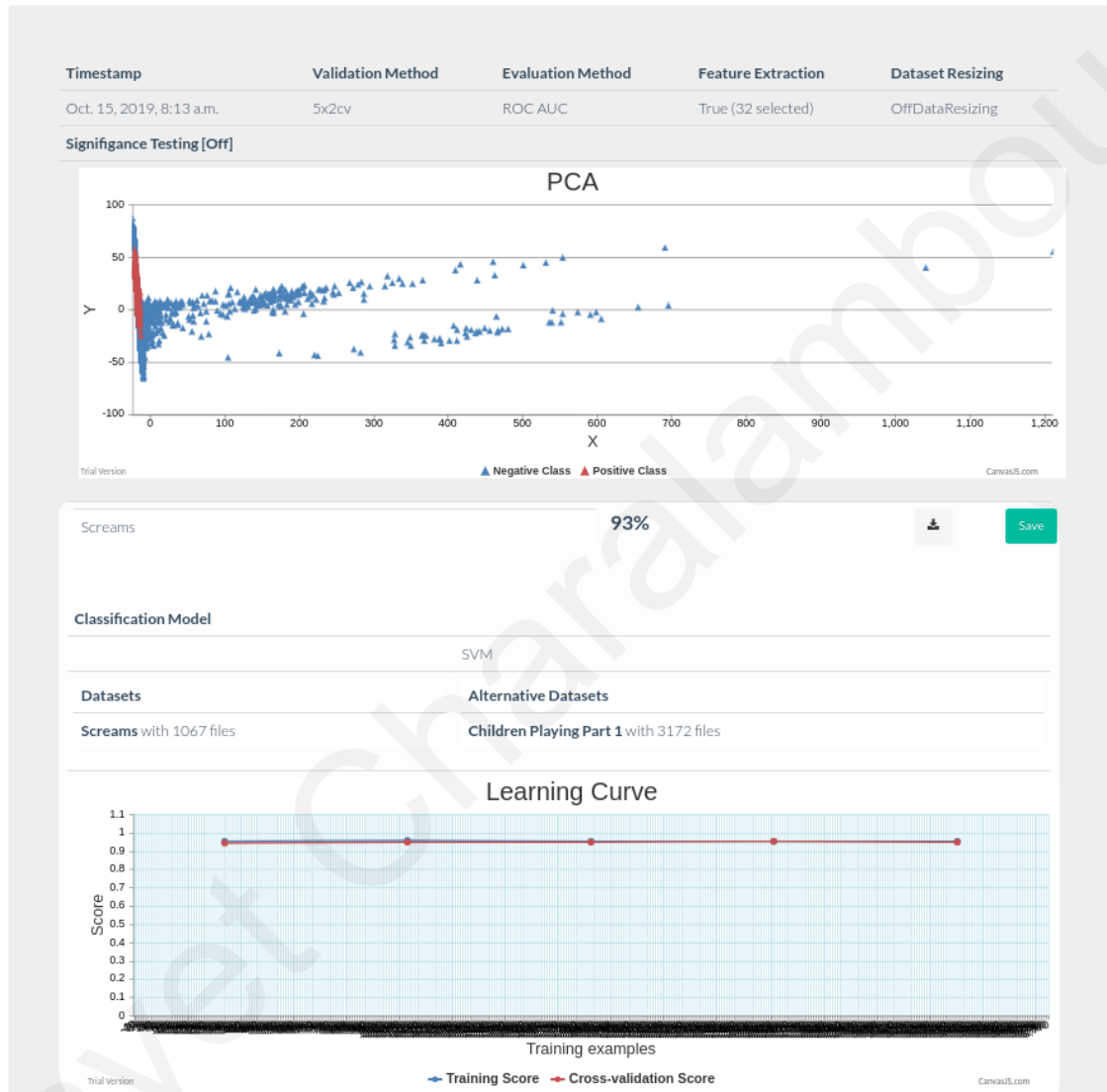


Figure 7.9: Detailed page of training request

Following the a request for the training of classification models, a user may observe the algorithm's capability in correctly classifying training and validation datasets. Trained classifiers with acceptable performance may be saved for use at a later stage with unseen data or to download the model for use through a different system (see Figure 7.9). As with the EDA panel, the Data Analysis tool supports for standardised interfaces which allow the integration of further classification algorithms as options.

Classification of Unseen data

Classification of unseen data is possible for trained models previously saved from the training results page. The user may select the relevant model from the drop-down list in the post analysis panel. As with the already discussed procedure for data import, a user may select a number of CSV or multimedia files for audio extraction (see Figure 7.10).

Create a new analysis request

Trained Models	
Random Forests PosEvent: Screams NegEvent: Gunshot Children_playing Perf:0.98	
Positive Event	screams
Negative Event	gunshot,children_playing
Classification Model	Random Forests
Evaluation	0.978
Training Datasets	Screams, Gunshots V2, Gunshots V1, Children Playing Part 1

Submit

Import data

Audio/Video CSV

Select files

The platform accepts audio and video files for analysis.

Selected files:

Figure 7.10: Classification of unseen data

For each uploaded file, data are sequentially read and fed as input to the indicated classifier. For audio classification, the uploaded files are first transcoded and split into blocks from which features are extracted. Data generated by each file are classified individually to allow visualisation of the attained results (see Figure 7.11).

When classification is completed, a new entry in the list of results is added in the table at the bottom of the page. The detailed view of each classification request visualises an accumulative view of all labeled data in a PCA plot at the top of the page (see Figure 7.12), followed by the results of classification for each file (see Figure 7.13). Classified samples are coloured to shades of blue and orange depending on which class they have been assigned; the shade of the colour is determined based on the classification confidence.

For each processed file, a single line colour-coded file-map provides a quick overview of the content of the file. Additionally, the classification label and confidence are reported for each sample in the form of a list.

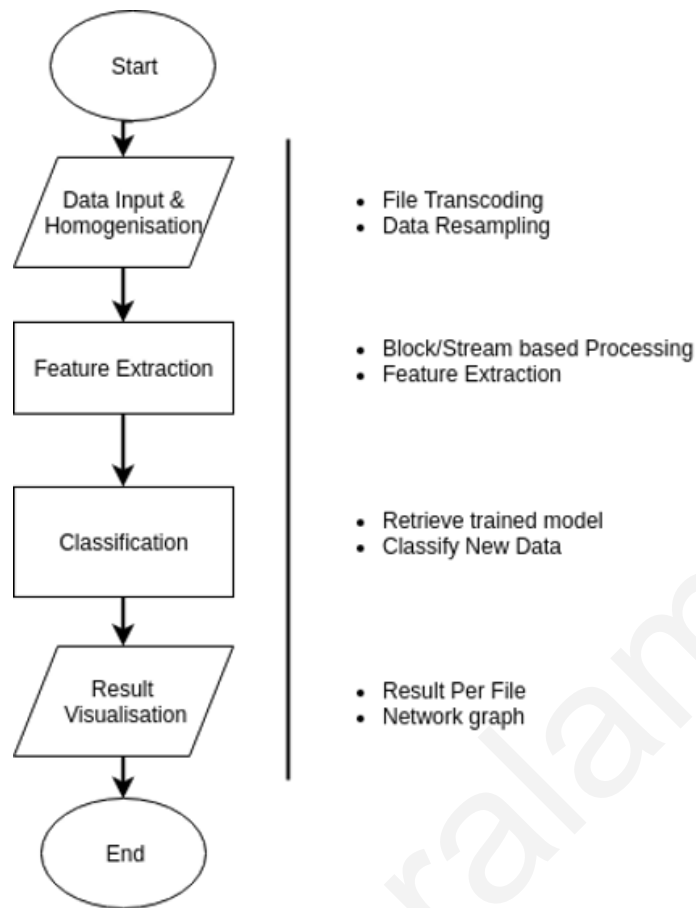


Figure 7.11: Data analysis flow through the data analysis suite.

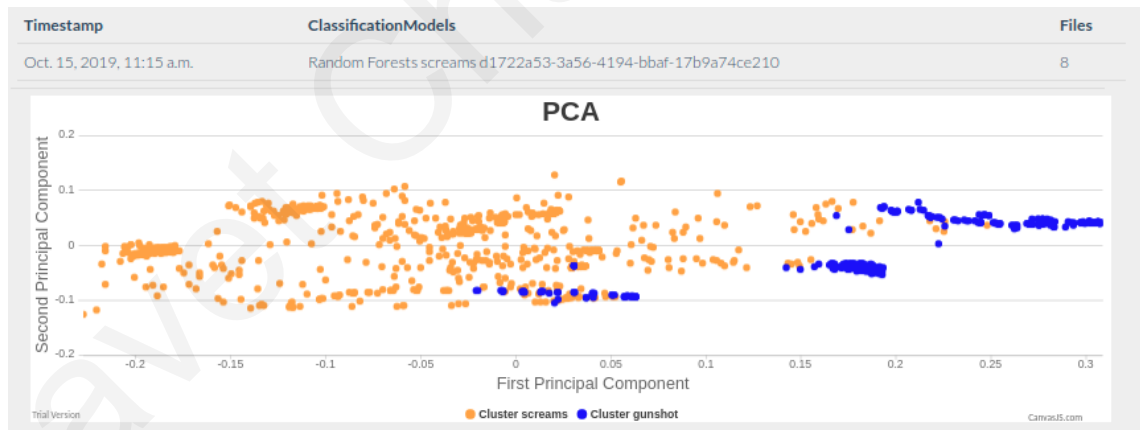


Figure 7.12: Data analysis flow through the data analysis suite.

7.5 Future Developments & Application

The demand for the development of data analysis tools like the Data Analysis Suite, is apparent even with the existence of numerous data analysis and analytics platforms and emerged from experience in several EU funded projects that collaborate closely with end users.

The current work has been applied in the fields of archaeology and acoustic event



Figure 7.13: Snippet of results for a file

detection for security purposes. Even though the two application domains do not seem to share much in common, they both impose specific restrictions in as far as data recording and sharing are concerned. Data in both domains are abundant and an easy to use end to end solution for exploratory data analysis and classification is of interest. The current tool has emerged based on the methodology presented in Chapter 4 applied in archaeological context within the framework of the Marie-Curie ITN project NARNIA and in the FP7 P-REACT and H2020 ASGARD projects for the training of acoustic event detection classifiers.

Future development plans for the tool are support for multi-class classification problems for the incorporation of confusion matrices in classification results for inter-class relationship analysis, as well as the the addition of the option to train classification models with parametrisation.

Chapter 8

Conclusions & Impact

Machine learning algorithms have many applications and are being deployed in interesting ways. It has become increasingly ubiquitous with more and more applications even in the most unlikely areas. The use of ML in so many domains and applications has led to an emerging need of application domain experts to, not only familiarise themselves with the basic principles of data analysis, but also to gain substantial comprehension on data sampling, transformation and analysis practices.

The decision making process followed by the human brain is highly complex and usually utilizes past knowledge, experiences and stimuli from the environment. The development of algorithms that mitigate part of this decision making process is challenging due to the many involved parameters; one of which is also the way we perceive data.

Machine learning techniques and statistical analysis can be very useful if used appropriately. Classification aims at identifying to which element of a set of categories, a new uncategorised artifact belongs, on the basis of a training set of artifacts the class/type of which is known. Classification algorithms are being developed under a set of assumptions which in practice may not always realistic, however their deployment needs to comply with the imposed restrictions and also assure that the algorithm's objectives are in accordance to the problem's underlying mechanisms; ensuring no conflicting constraints.

Multiple factors contribute to the introduction of noise or uncertainty in data measurements; these might be both intrinsic and extrinsic. The sole categorisation of data with classification methods requires the learning of robust classifiers with validated performance. The utilisation of prior knowledge or actual models of uncertainty may

significantly improve the accuracy of classification; however these are not available in realistic conditions.

The sampling procedure of – the under analysis – artifacts to their quantitative or qualitative representation generally introduces uncertainty emerging from sampling instrumentation, feature extraction and selection procedures as well as the selection and parametrisation of classification methods; uncertainty is sometimes also introduced when artifacts are imperfect even in their original form (i.e damaged or tempered tangible artifacts, insufficient sampling or inappropriate recording equipment for intangible artifacts). Considering, the complexity and involved factors, it is rather challenging to train robust classification models with validated performance without coding or scripting involved.

Therefore, this thesis investigated the classification problem and its contributing factors through different perspectives and proposed a methodology for the robust classification of heterogeneous and scarce data under uncertainty. Aspects relevant to the use of distance metrics, the heterogeneity in size and composition within data, the contribution of extracted features, the evaluation procedure and the bias due to overfitting were analysed with respect to their impact. Through this analysis a design based on well established and standardised methods is proposed to support a two fold purpose:

- To allow the comparative analysis between candidate classification algorithms on a specific task measuring the classification performance of the trained models and also its robustness upon prediction;
- To reveal inter-class relationships between categories of artifacts and also between classes and mis-classified samples.

The suggested design has been validated through application on two vastly different domains through independent case studies. The Case Study I involved the analysis of compositional archaeological data with uncertainties and practices concerning the multi-class classification problem with the use of annotated data from an archaeologist expert. The archaeological process in the categorisation of artifacts in the appropriate fabric relies heavily on macroscopic, microscopic and compositional evaluations as well as on prior knowledge from past analyses. Even though classification is applied successfully in many domains, no standardised methodology has been so far reported for the effective classification of archaeological artifacts. This is emphasised by the

fact that, classification on chemical compositional data is applied at the end of this thorough process for the archaeologist to validate whether the findings of their analysis match the compositional structure of the artifact.

Therefore, the characteristics and constraints of the analysis of chemical compositional data were analysed and the proposed methodology has been deployed with the appropriate configuration to allow for sound statistical conclusions. An experimental design for the training of robust classifiers for archaeological ceramic compositional data was presented in Chapter 5 where the robustness of classification on complicated and highly overlapping data was successfully validated with the use of statistically valid methods for the Simplex space where chemical compositional data lay. Additionally, the implementation or re-sampling allowed for the categorisation of samples marked as outliers by experts; a task that could not be solved with certainty beforehand due to the absence of discriminating petrology in the ceramic thin sections. Finally, post analysis on mis-classification pattered resulted in conclusions with respect to inter-class relationships that revealed links with respect to technological similarities and chronological evolution.

Through this study it was demonstrated that robust classification may assist the archaeological process in ways that no other currently deployed form of analysis could. Categorisation based on subset groups of chemical elements and the identification of inter-class relationships are impactful to the archaeological community. Additionally, the utilisation of artifacts that were marked as outliers, due to their inconsistent findings from other types of analysis, revealed links with other classes. Classification in the archaeological process on ceramic data, through the implementation of the proposed methodology, assisted in the recognition and validation of compositional patterns, and the identification of possible categorisation mistakes.

Considering that generalisation of the methodology to other domains is also important. The designed methodology as part of Case Study II, was also deployed in the security and surveillance domain for the implementation of robust classification models in the field of acoustic event detection. The auditory scene in an environment is highly dynamic – and unpredictable – making the training of robust classifiers a real challenge. Acoustic events may adhere to very diverse characteristics with the perception of sound significantly impacted by the surrounding environment. Since models trained for surveillance aim use in open and unrestricted environments, effective operation of the systems – serving the already defined specification – require analysis of the data

with multiple techniques where each one examines different aspects of the sample.

The aim of this study was to train robust classifiers for the detection of gunshots, glass breaking and screaming incidents for use on resource constraint embedded systems. The trained models were designed to be deployed as part of an ethical audio surveillance system developed for the cost effective and real-time detection of auditory events of interest.

The desired solution should implement classifiers for first level of analysis in the resource constrained device and second level of analysis on a private cloud. The selection of audio analytics models for the first and second levels of analysis followed a comparative study in which re-sampling and boosting, which are integral parts of the methodology, allowed generating new datasets with a balanced number of samples for uniform class distribution. The selection of the preferred classification method in this case study, not only depended on the classification performance, but also on the required processing time.

The objectives of this Case Study were to first evaluate the performance of a number of algorithms in successfully performing audio event detection and secondly their evaluation in terms of time complexity as the solution aided deployment on a low cost embedded system. After experimentation, a series of binary classifiers were trained and deployed in cascaded form for the first layer of analysis, For the second level of analysis, samples are passed through a series of binary classifiers one for each acoustic event of interest, the classification outcome emerges as a result of majority voting.

Both case studies validated that a configurable but standardised methodology for classification may allow the learning of robust models for consistent analysis outcome. Even though the analysis of chemical compositional data and sound waves do not have much in common, they are impacted by similar challenges. As a result classification analysis on data should be applied only after sufficient comprehension of the respective domain or through the use of implemented tools that implement these structures internally.

8.1 Future Work

Research on the objectives of this thesis and the validation of the presented methodology through two case studies, generated a number of additional research questions for future investigation.

The analysis of chemical compositional data and sound waves revealed that the appropriate distance metrics and techniques should be used for a statistically valid classification outcome. In the case of chemical compositional data, the use of the Aitchinson distance is advised, while for the analysis of frequency domain features the use of Manhattan distance or χ^2 are preferred. The emergence however of methods that require the combination of features that have not been extracted under the same basis (i.e. combination of time and frequency domain features or the combination of linearly and non-linearly calculated measurements); further research is required in this respect.

Findings in the analysis of inter-class relationships led us to believe that since different algorithms are better in exploiting specific characteristics of the data, the "aggregation" of their produced results would benefit the expert's analysis; this is possible with the use of ensemble methods and majority voting techniques. Interestingly, since the operation of a classification algorithm is determined based on its underlying mathematical definitions, complementarity analysis to determine whether the decision making process of classifiers is based on different criteria. Knowledge on the complementarity of classifiers is expected to increase the individual classifiers performance with the use of classification fusion approaches.

The implementation of the Data Analysis Suite tool, presented in Chapter 7 allows the rapid training and deployment of robust classifiers. Experimentation with the tool enables researchers to draw interesting findings with respect to characteristics of their data, and the impact of algorithms with respect to the result. The standardisation of interfaces to support the addition of further functionalities expand its configurability. However, not all methodology steps have been integrated in the tool to minimise the training time. Due to the interest drawn by end-users and its ease-of-use, boosting, multi-class classification and the automated detection of inter-class relationships will be integrated. Doing so will allow reproducibility of the presented case studies – and others – without the need for scripting or coding.

Elisavet Charalambous

Bibliography

- [1] S. Abe, Support vector machines for pattern classification. Springer, 2005, vol. 2.
- [2] A. Acton, Issues in Environmental Research and Application: 2013 Edition. ScholarlyEditions, 2013. [Online]. Available: <http://books.google.com/books?id=YGnl6wBSfjYC\&pgis=1>
- [3] E. Adebayo, I. Jibrin Enejo, and I. Muhammed Lawal, “Awareness and Level of Competency of Academic Social Networking Sites for Research among Post-graduate Students in South-West, Nigeria,” *International Journal of Science and Research Methodology (IJSRM)*, vol. 14, no. 2, pp. 17–28, 2019.
- [4] C. C. Aggarwal and P. S. Yu, “Outlier detection with uncertain data,” in *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 2008, pp. 483–493.
- [5] J. Aitchison, C. Barceló-Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn, “Logratio analysis and compositional distance,” *Mathematical Geology*, vol. 32, no. 3, pp. 271–275, 2000.
- [6] J. Aitchison, “The statistical analysis of compositional data,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.
- [7] J. Aitchison, C. Barceló-Vidal, and V. Pawlowsky-Glahn, “Some comments on compositional data analysis in archaeometry, in particular the fallacies in tangri and wright’s dismissal of logratio analysis,” *Archaeometry*, vol. 44, no. 2, pp. 295–304, 2002.
- [8] J. Aitchison and J. J. Egozcue, “Compositional data analysis: where are we and where should we be heading?” *Mathematical Geology*, vol. 37, no. 7, pp. 829–850, 2005.
- [9] E. Alpaydin, “Introduction to machine learning. sl,” 2010.
- [10] E. Alpaydm, “Combined 5×2 cv f test for comparing supervised classification learning algorithms,” *Neural computation*, vol. 11, no. 8, pp. 1885–1892, 1999.
- [11] S. Arafat, M. Dohrmann, and M. Skubic, “Classification of coronary artery disease stress ecgs using uncertainty modeling,” in *2005 ICSC Congress on Computational Intelligence Methods and Applications*. IEEE, 2005, pp. 4–pp.
- [12] N. Arivazagan and M. RAMULA, “Whatsapp as an interactive teaching tool in surgery for undergraduate medical students,” *IOSR Journal of Dental and Medical Sciences*, vol. 18, no. 7, pp. 75–78, 2019.

- [13] R. Artstein and M. Poesio, “Inter-coder agreement for computational linguistics,” *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.
- [14] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, “Audio based event detection for multimedia surveillance,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5. IEEE, 2006, pp. V–V.
- [15] M. Baillie and J. M. Jose, “Audio-based event detection for sports video,” in *International Conference on Image and Video Retrieval*. Springer, 2003, pp. 300–309.
- [16] J. A. Barcelo, *Computational Intelligence in Archaeology*. Information Science Reference, 2009.
- [17] J. A. Barceló and J. A. Barcelo, *Computational intelligence in archaeology*. Information Science Reference, 2009.
- [18] A. G. Barto and R. Sutton, “Introduction to reinforcement learning,” 1997.
- [19] M. J. Baxter, *Exploratory multivariate analysis in archaeology*. Edinburgh University Press, 1994.
- [20] —, “A review of supervised and unsupervised pattern recognition in archaeometry,” *Archaeometry*, vol. 48, no. 4, pp. 671–694, 2006.
- [21] M. Ben-Bassat, “Use of distance measures information measures and error bounds in feature selection,” in *The Handbook of Statistics, II*. North Holland, 1981.
- [22] A. Ben-Tal, S. Bhadra, C. Bhattacharyya, and A. Nemirovski, “Efficient methods for robust classification under uncertainty in kernel matrices,” *Journal of Machine Learning Research*, vol. 13, no. Oct, pp. 2923–2954, 2012.
- [23] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*. Princeton University Press, 2009, vol. 28.
- [24] A. Bendale and T. Boulton, “Towards open world recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1893–1902.
- [25] A. Bendale and T. E. Boulton, “Towards open set deep networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.
- [26] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. D. Zhuo, “Robust classification,” *INFORMS Journal on Optimization*, vol. 1, no. 1, pp. 2–34, 2018.
- [27] J. C. Bezdek and R. J. Hathaway, “Vat: A tool for visual assessment of (cluster) tendency,” in *Neural Networks, 2002. IJCNN’02. Proceedings of the 2002 International Joint Conference on*, vol. 3. IEEE, 2002, pp. 2225–2230.
- [28] J. Bi and T. Zhang, “Support vector classification with input data uncertainty,” in *Advances in neural information processing systems*, 2005, pp. 161–168.
- [29] C. M. Bishop, “Pattern recognition,” *Machine Learning*, 2006.

- [30] E. O. Biu, M. T. Nwakuya, and N. Wonu, "Detection of non-normality in data sets and comparison between different normality tests," *Asian Journal of Probability and Statistics*, pp. 1–20, 2019.
- [31] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang et al., "End to end learning for self-driving cars," arXiv preprint arXiv:1604.07316, 2016.
- [32] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.
- [33] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.
- [34] D. Boyce, A. Farhi, and R. Weischedel, *Optimal Subset Selection: Multiple Regression, Interdependence and Optimal ...* - David Boyce, A. Farhi, R. Weischedel - Google Books. Springer-Verlag Berlin Heidelberg, 2013.
- [35] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [36] A. H. Briggs, M. C. Weinstein, E. A. Fenwick, J. Karnon, M. J. Sculpher, A. D. Paltiel, I.-S. M. G. R. P. T. Force et al., "Model parameter estimation and uncertainty: a report of the ispor-smdm modeling good research practices task force-6," *Value in Health*, vol. 15, no. 6, pp. 835–842, 2012.
- [37] H. Broekhuizen, C. G. Groothuis-Oudshoorn, J. A. van Til, J. M. Hummel, and M. J. IJzerman, "A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions," *Pharmacoeconomics*, vol. 33, no. 5, pp. 445–455, 2015.
- [38] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24–38, 2006.
- [39] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.
- [40] J. Bryan-Jones, "On an error in" instructions to authors"," *HortScience*, vol. 18, pp. 279–282, 1983.
- [41] J. Buxeda, "Revisiting the compositional data. Some fundamental questions and new prospects in Archaeometry and Archaeology," May 2008. [Online]. Available: <http://dugi-doc.udg.edu/handle/10256/749>
- [42] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 3, pp. 1026–1039, 2006.
- [43] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," in *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, vol. 3. IEEE, 2003, pp. III–37.
- [44] G. C. Cawley, N. L. Talbot, and M. Girolami, "Sparse multinomial logistic regression via bayesian l1 regularisation," in *Advances in neural information processing systems*, 2007, pp. 209–216.

- [45] O. Chapelle, B. Scholkopf, and A. Zien, “Semi-supervised learning (chappelle, o. et al., eds.; 2006)[book reviews],” *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.
- [46] E. Charalambous, M. Dikomitou-Eliadou, G. M. Milis, G. Mitsis, and D. G. Eliades, “An experimental design for the classification of archaeological ceramic data from cyprus, and the tracing of inter-class relationships,” *Journal of Archaeological Science: Reports*, 2015.
- [47] E. Charalambous, N. Efstathiou, and N. Koutras, “A Cost Effective Solution for Audio Surveillance Using Embedded Devices as Part of a Cloud Infrastructure,” in *International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE’2015)*, A. Rumen D., Ed. Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences, 2015, pp. 21–31. [Online]. Available: http://conference.ott-iict.bas.bg/wp-content/uploads/2015/12/BdKCSE2015{_}Proseedings1.pdf
- [48] F. Chayes, “On ratio correlation in petrography,” *The Journal of Geology*, pp. 239–254, 1949.
- [49] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “Deepdriving: Learning affordance for direct perception in autonomous driving,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2722–2730.
- [50] R. Chen and I. Paschalidis, “Outlier detection using robust optimization with uncertainty sets constructed from risk measures,” *ACM SIGMETRICS Performance Evaluation Review*, vol. 45, no. 3, pp. 174–179, 2018.
- [51] V. Chew, “Uses and abuses of duncan’s multiple range test.” in *Proceedings of the Florida State Horticultural Society*, 1977.
- [52] S.-B. Cho and J. H. Kim, “Combining multiple neural networks by fuzzy integral for robust classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 2, pp. 380–384, 1995.
- [53] S.-S. Choi, S.-H. Cha, and C. C. Tappert, “A survey of binary similarity and distance measures,” *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43–48, 2010.
- [54] S. Chu, S. Narayanan, and C.-C. J. Kuo, “Environmental sound recognition with time–frequency audio features,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.
- [55] A. Clare and R. D. King, “Knowledge discovery in multi-label phenotype data,” in *Principles of data mining and knowledge discovery*. Springer, 2001, pp. 42–53.
- [56] C. Clavel, T. Ehrette, and G. Richard, “Events detection for an audio-based surveillance system,” in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 1306–1309.
- [57] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [58] H. Coolican, *Research methods and statistics in psychology*. Hodder & Stoughton Educational, 1990.

- [59] W. K. Cornwell, D. W. Schwilk, and D. D. Ackerly, "A trait-based test for habitat filtering: convex hull volume," *Ecology*, vol. 87, no. 6, pp. 1465–1471, 2006.
- [60] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [61] Z. Dao and S. Chen, "Kernel-based fuzzy and possibilistic c-means clustering," in *Proceedings of the International Conference Artificial Neural Network*, 2003.
- [62] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.
- [63] M. D. de Lima, J. d. O. R. e Lima, and R. M. Barbosa, "Medical data set classification using a new feature selection algorithm combined with twin-bounded support vector machine," *Medical & Biological Engineering & Computing*, pp. 1–10, 2020.
- [64] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le et al., "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.
- [65] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice hall, 1982.
- [66] T. G. Dietterich, "Statistical tests for comparing supervised classification learning algorithms," *Oregon State University Technical Report*, vol. 1, pp. 1–24, 1996.
- [67] —, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [68] —, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [69] —, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.
- [70] M. Dikomitou, "Ceramic production, distribution, and social interaction. an analytical approach to the study of early and middle bronze age pottery from cyprus." Ph.D. dissertation, University College London Institute of Archaeology, 2012.
- [71] —, "Ceramic production, distribution, and social interaction. an analytical approach to the study of early and middle bronze age pottery from cyprus." Ph.D. dissertation, UCL (University College London), 2012.
- [72] M. Dikomitou-Eliadou, "Interactive communities at the dawn of the cypriot bronze age: an interdisciplinary approach to philia phase ceramic variability," *JRB Stewart: An Archaeological Legacy on Cyprus*; Knapp, AB, Webb, JM, McCarthy, A., Eds, pp. 23–31, 2013.
- [73] —, "Rescaling perspectives: local and island-wide ceramic production in early and middle bronze age cyprus," *Structure, Measurement and Meaning. Studies on Prehistoric Cyprus in Honour of David Frankel*, vol. 14, pp. 199–211, 2014.

- [74] J. Dolata, H.-J. Mucha, and H.-G. Bartel., “Uncovering the internal structure of the roman brick and tile making inuncovering the internal structure of the roman brick and tile making in frankfurt-nied by cluster validation,” in *Advances in data analysis*. Springer Berlin Heidelberg, 2007, pp. 663–670.
- [75] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 1998, vol. 326.
- [76] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
- [77] S. Dumais and H. Chen, “Hierarchical classification of web content,” in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000, pp. 256–263.
- [78] J. G. Dy and C. E. Brodley, “Feature selection for unsupervised learning,” *Journal of machine learning research*, vol. 5, no. Aug, pp. 845–889, 2004.
- [79] R. Eastell, S. L. Cedel, H. W. Wahner, B. L. Riggs, and L. J. Melton III, “Classification of vertebral fractures,” *Journal of Bone and Mineral Research*, vol. 6, no. 3, pp. 207–215, 1991.
- [80] B. Efron and R. Tibshirani, “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy,” *Statistical science*, pp. 54–75, 1986.
- [81] E. Elhamifar and R. Vidal, “Robust classification using structured sparse representation,” in *CVPR 2011*. IEEE, 2011, pp. 1873–1879.
- [82] A. Elisseeff and J. Weston, “A kernel method for multi-labelled classification,” in *Advances in neural information processing systems*, 2001, pp. 681–687.
- [83] D. Ellis, “Chroma feature analysis and synthesis,” *Resources of Laboratory for the Recognition and Organization of Speech and Audio-LabROSA*, 2007.
- [84] D. P. Ellis, “Prediction-driven computational auditory scene analysis,” Ph.D. dissertation, Columbia University, 1996.
- [85] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, “Audio-based context recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2005.
- [86] D. G. Everitt and G. Dunn, “An introduction to mathematical taxonomy,” Cambridge: Cambridge University, 1982.
- [87] Y. Fan and R. Wang, “An image retrieval method using dct features,” *Journal of Computer Science and Technology*, vol. 17, no. 6, pp. 865–873, 2002.
- [88] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, “Comparison of evaluation metrics in classification applications with imbalanced datasets,” in *2008 Seventh International Conference on Machine Learning and Applications*. IEEE, 2008, pp. 777–782.
- [89] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

- [90] P. Fermo, E. Delnevo, M. Lasagni, S. Polla, and M. de Vos, "Application of chemical and chemometric analytical techniques to the study of ancient ceramics from dougga (tunisia)," *Microchemical journal*, vol. 88, no. 2, pp. 150–159, 2008.
- [91] A. D. Forbes, "Classification-algorithm evaluation: Five performance measures based on confusion matrices," *Journal of Clinical Monitoring*, vol. 11, no. 3, pp. 189–206, 1995.
- [92] G. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Transactions on Information theory*, vol. 18, no. 3, pp. 363–378, 1972.
- [93] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [94] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.
- [95] G. Ganpath, "What are some interesting possible applications of machine learning?" Jan. 2015. [Online]. Available: <https://www.quora.com/What-are-some-interesting-possible-applications-of-machine-learning>
- [96] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.
- [97] S. Geisser, *Predictive inference*. CRC press, 1993, vol. 55.
- [98] D. Gerhard, *Audio signal classification: History and current techniques*. Cite-seer, 2003.
- [99] G. Geser, "Open access and open data in archaeology: Report on the ariadne session," in *European Association of Archaeologists 20th Annual Meeting*. Salzburg Research, 2014. [Online]. Available: http://www.ariadne-infrastructure.eu/ger/content/download/4345/25141/file/ARIADNE_EAA2014_OpenAccess_session_Report.pdf
- [100] J. Gill, "Current status of multiple comparisons of means in designed experiments," *Journal of Dairy Science*, vol. 56, no. 8, pp. 973–977, 1973.
- [101] S. Gopal and Y. Yang, "Multilabel classification with meta-level features," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 315–322.
- [102] N. J. Gotelli, "Null model analysis of species co-occurrence patterns," *Ecology*, vol. 81, no. 9, pp. 2606–2621, 2000.
- [103] N. Gotelli and G. Graves, *Null models in ecology*. Smithsonian Institution Press, 1996. [Online]. Available: <https://books.google.com.cy/books?id=fGnwAAAAMAAJ>
- [104] R. Gour, "Working structure of django mtv architecture - towards data science." [Online]. Available: <https://towardsdatascience.com/working-structure-of-django-mtv-architecture-a741c8c64082>

- [105] P. C. R. Group et al., “The study of prehistoric pottery: general policies and guidelines for analysis and publication,” *Occasional Papers*, no. 1, 2010.
- [106] M. Guillaumin, J. Verbeek, and C. Schmid, “Multimodal semi-supervised learning for image classification,” in *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society, 2010, pp. 902–909.
- [107] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [108] A. Halder and T. Mandal, “Etiological factors of post menopausal bleeding in a tertiary care hospital.” *Hypertension*, vol. 60, no. 3, pp. 201–204, 2010.
- [109] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, “Quality scheme assessment in the clustering process,” in *Principles of Data Mining and Knowledge Discovery*. Springer, 2000, pp. 265–276.
- [110] L. K. Hansen and P. Salamon, “Neural network ensembles,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 993–1001, 1990.
- [111] A. Harma, M. F. McKinney, and J. Skowronek, “Automatic surveillance of the acoustic activity in our living environment,” in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 4–pp.
- [112] C. Harte, M. Sandler, and M. Gasser, “Detecting harmonic change in musical audio,” in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. ACM, 2006, pp. 21–26.
- [113] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [114] T. Hastie and R. Tibshirani, “Discriminant adaptive nearest neighbor classification,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 6, pp. 607–616, 1996.
- [115] T. C. Havens, J. C. Bezdek, J. M. Keller, and M. Popescu, “Clustering in ordered dissimilarity data,” *International Journal of Intelligent Systems*, vol. 24, no. 5, pp. 504–528, May 2009. [Online]. Available: <http://dx.doi.org/10.1002/int.20344>
- [116] X. He, O. King, W.-Y. Ma, M. Li, and H.-J. Zhang, “Learning a semantic space from user’s relevance feedback for image retrieval,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 1, pp. 39–48, 2003.
- [117] T. Heittola and A. Klapuri, “Tut acoustic event detection system 2007,” in *multimodal technologies for perception of humans*. Springer, 2007, pp. 364–370.
- [118] G. Heitz, S. Gould, A. Saxena, and D. Koller, “Cascaded classification models: Combining models for holistic scene understanding,” in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 641–648. [Online]. Available: <http://papers.nips.cc/paper/3472-cascaded-classification-models-combining-models-for-holistic-scene-understanding.pdf>

- [119] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," arXiv preprint arXiv:1610.02136, 2016.
- [120] H. Hermansky, "Mel cepstrum, deltas, double-deltas,...-what else is new," Proc. Robust Methods for Speech Recognition in Adverse Condition, 1999.
- [121] N. T. Hobbs and R. Hilborn, "Alternatives to statistical hypothesis testing in ecology: a guide to self teaching," *Ecological Applications*, vol. 16, no. 1, pp. 5–19, 2006.
- [122] D. Hoiem, Y. Ke, and R. Sukthankar, "Solar: Sound object localization and retrieval in complex audio environments," in Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 5. IEEE, 2005, pp. v–429.
- [123] A. Holovaty and J. Kaplan-Moss, *The definitive guide to Django: Web development done right*. Apress, 2009.
- [124] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015.
- [125] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.
- [126] M. Huang, Z. Xia, H. Wang, Q. Zeng, and Q. Wang, "The range of the value for the fuzzifier of the fuzzy c-means algorithm," *Pattern Recognition Letters*, vol. 33, p. 2280?2284, 2012.
- [127] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification* 2.1, pp. 193–218, 1985.
- [128] Y. Ikhari, V. Gade, S. Patil, and J. Gade, "The Influence of Various Irrigants on The Accuracy of Third Generation Apex Locator And Fifth Generation Apex Locators In Locating Simulated Root Perforation: An In Vitro Study," *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS) e-ISSN*, vol. 18, pp. 86–91, 2019. [Online]. Available: www.iosrjournals.org
- [129] V. Ilakovac et al., "Statistical hypothesis testing and some pitfalls," *Biochemia Medica*, vol. 19, no. 1, pp. 10–16, 2009.
- [130] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.patrec.2009.09.011>
- [131] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
- [132] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

- [133] J. Jenness and J. J. Wynne, “Cohen’s kappa and classification table metrics 2.0: An arcview 3. x extension for accuracy assessment of spatially explicit models,” Open-File Report OF 2005-1363. Flagstaff, AZ: US Geological Survey, Southwest Biological Science Center. 86 p, 2005.
- [134] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, “Music type classification by spectral contrast feature,” in Proceedings. IEEE International Conference on Multimedia and Expo, vol. 1. IEEE, 2002, pp. 113–116.
- [135] H. Jiang, B. Kim, M. Guan, and M. Gupta, “To trust or not to trust a classifier,” in Advances in neural information processing systems, 2018, pp. 5541–5552.
- [136] G. H. John, R. Kohavi, and K. Pfleger, “Irrelevant features and the subset selection problem,” in MACHINE LEARNING: PROCEEDINGS OF THE ELEVENTH INTERNATIONAL. Morgan Kaufmann, 1994, pp. 121–129.
- [137] D. H. Johnson, “The insignificance of statistical significance testing,” The journal of wildlife management, pp. 763–772, 1999.
- [138] S. Johnson and R. Berger, “On the status of statistics in phytopathology [journals].” *Phytopathology*, 1982.
- [139] D. Jones and N. Matloff, “Statistical hypothesis testing in biology: a contradiction in terms,” *Journal of Economic Entomology*, vol. 79, no. 5, pp. 1156–1160, 1986.
- [140] M. Kanevski, V. Timonin, and A. Pozdnukhov, *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press, 2009.
- [141] V. Kassianidou, M. Dikomitou-Eliadou, B. Κασσιανίδου, and M. Δικωμίτου-Ηλιάδου, *The NARNIA Project: Integrating approaches to ancient material studies*. NARNIA Project, 2014.
- [142] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda, “Maximal margin labeling for multi-topic text categorization,” in Advances in neural information processing systems, 2004, pp. 649–656.
- [143] S. W. Kembel, P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg, and C. O. Webb, “Picante: R tools for integrating phylogenies and ecology,” *Bioinformatics*, vol. 26, no. 11, pp. 1463–1464, 2010.
- [144] H. Kim, C. M. Park, and J. M. Goo, “Test-retest reproducibility of a deep learning-based automatic detection algorithm for the chest radiograph,” *European Radiology*, pp. 1–10, 2019.
- [145] T. Kohonen, “Learning vector quantization,” in *Self-organizing maps*. Springer, 1995, pp. 175–189.
- [146] D. Koller and M. Sahami, “Hierarchically classifying documents using very few words,” *Stanford InfoLab, Tech. Rep.*, 1997.
- [147] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” 2007.

- [148] B. Kowalski and C. Bender, "Pattern recognition. ii. linear and nonlinear methods for displaying chemical data," *Journal of the American Chemical Society*, vol. 95, no. 3, pp. 686–693, 1973.
- [149] B. Kowalski, T. Schatzki, and F. Stross, "Classification of archaeological artifacts by applying pattern recognition to trace element data," *Analytical Chemistry*, vol. 44, no. 13, pp. 2176–2180, 1972.
- [150] M. E. Kreye, Y. M. Goh, and L. B. Newnes, "Manifestation of uncertainty-a classification," in *DS 68-6: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 6: Design Information and Knowledge, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011, 2011*.
- [151] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [152] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013, vol. 26.
- [153] D. H. Kulkarni, "Computational statistics and predictive analysis in machine learning," *International Journal of Science and Research (IJSR)*, vol. 5, 2016.
- [154] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification." in *ICCV*, vol. 2, no. 1, 2005.
- [155] M. B. Kursa, W. R. Rudnicki et al., "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.
- [156] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li et al., *Applied linear statistical models*. McGraw-Hill Irwin Boston, 2005, vol. 5.
- [157] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, "Stability-based validation of clustering solutions," *Neural computation*, vol. 16, no. 6, pp. 1299–1323, 2004.
- [158] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," *arXiv preprint arXiv:1711.09325*, 2017.
- [159] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, 2018, pp. 7167–7177.
- [160] H. Liu and H. Motoda, *Computational methods of feature selection*. CRC Press, 2007.
- [161] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge & Data Engineering*, no. 4, pp. 491–502, 2005.
- [162] B. Logan et al., "Mel frequency cepstral coefficients for music modeling." in *IS-MIR*, vol. 270, 2000, pp. 1–11.
- [163] A. Lopez-Molinero, A. Castro, J. Pino, J. Perez-Arantegui, and J. R. Castillo, "Classification of ancient roman glazed ceramics using the neural network of self-organizing maps," *Fresenius' journal of analytical chemistry*, vol. 367, no. 6, pp. 586–589, 2000.

- [164] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, 2002.
- [165] L. Madden, J. Knoke, and R. Louie, "Considerations for the use of multiple comparison procedures in phytopathological investigations," *Phytopathology*, vol. 72, pp. 1015–1017, 1982.
- [166] C. L. Mallows, "Some comments on cp," *Technometrics*, vol. 42, no. 1, pp. 87–94, 2000.
- [167] C. D. Manning, R. Prabhakar, and S. Hinrich, *Introduction to information retrieval*. Cambridge: Cambridge University Press, 2008, vol. 1.
- [168] B. McCune, J. Grace, and D. Urban, "Distance measures," *Analysis of Ecological Communities*. Glenden Beach, OR: MJM, pp. 45–57, 2002.
- [169] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
- [170] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European Signal Processing Conference*. IEEE, 2010, pp. 1267–1271.
- [171] A. Miller, *Subset selection in regression*. Chapman and Hall/CRC, 2002.
- [172] B. Mirkin, *Mathematical Classification and Clustering (Nonconvex Optimization and Its Applications)*. Kluwer Academic Press, 1996.
- [173] P. Mirti, R. Aruga, L. Appolonia, A. Casoli, and M. Oddone, "On the role of major, minor and trace elements in provenancing ceramic material. a case study: Roman terra sigillata from augusta praetoria," *Fresenius' journal of analytical chemistry*, vol. 348, no. 5-6, pp. 396–401, 1994.
- [174] M. Mirti, P., Aruga, R., Appolonia, L., Casoli, A., Oddone, "On the role of major, minor and trace elements in provenancing ceramic material. A case study: Roman terra sigillata from Augusta Praetoria," *Fresenius' Journal of Analytical Chemistry*, no. 348, pp. 396–401, 1994.
- [175] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 301–312, 2002.
- [176] J. C. Montero-Serrano, J. Palarea-Albaladejo, J. A. Martín-Fernández, M. Martínez-Santana, and J. V. Gutiérrez-Martín, "Sedimentary chemofacies characterization by means of multivariate analysis," *Sedimentary Geology*, vol. 228, no. 3, pp. 218–228, 2010.
- [177] B. M. Moret, "Decision trees and diagrams," *ACM Computing Surveys (CSUR)*, vol. 14, no. 4, pp. 593–623, 1982.
- [178] H. Müller, T. Pun, and D. Squire, "Learning from user behavior in image retrieval: Application of market basket analysis," *International Journal of Computer Vision*, vol. 56, no. 1-2, pp. 65–77, 2004.

- [179] C. Munita, A. Nascimento, S. Schreiber, S. Luna, and P. Oliveira, “Chemical study of some ceramics from brazilian northeast,” *Journal of Radioanalytical and Nuclear Chemistry*, vol. 259, no. 2, pp. 305–309, 2004.
- [180] K. Murphy, *Machine Learning: A Probabilistic Perspective*, ser. Adaptive computation and machine learning series. MIT Press, 2012. [Online]. Available: <https://books.google.com.cy/books?id=NZP6AQAAQBAJ>
- [181] G. Musumarra, M. Stella, M. Matteini, and M. Rizzi, “Multiariate characterization, using the simca method, of mortars from two frescoes in chiaravalle abbey,” *Thermochimica acta*, vol. 269, pp. 797–807, 1995.
- [182] M. R. Naphade, A. Garg, and T. S. Huang, “Duration dependent input output markov models for audio-visual event detection,” in *2001 IEEE International Conference on Multimedia and Expo, ICME 2001*. IEEE Computer Society, 2001, pp. 253–256.
- [183] P. M. Narendra and K. Fukunaga, “A branch and bound algorithm for feature subset selection,” *IEEE Transactions on computers*, no. 9, pp. 917–922, 1977.
- [184] S. Narkhede, “Understanding AUC - ROC Curve - Towards Data Science.” [Online]. Available: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- [185] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér, “Consistent feature selection for pattern recognition in polynomial time,” *Journal of Machine Learning Research*, vol. 8, no. Mar, pp. 589–612, 2007.
- [186] N. R. Pal and J. C. Bezdek, “Measuring fuzzy uncertainty,” *IEEE Transactions on Fuzzy Systems*, vol. 2, no. 2, pp. 107–118, 1994.
- [187] —, “Quantifying different facets of fuzzy uncertainty,” in *Fundamentals of Fuzzy Sets*. Springer, 2000, pp. 459–480.
- [188] D. Pascual, F. Pla, and J. S. Sánchez, “Cluster validation using information stability measures,” *Pattern Recognition Letters*, vol. 31, no. 6, pp. 454–461, 2010.
- [189] —, “Cluster validation using information stability measures,” *Pattern Recognition Letters*, vol. 31, no. 6, pp. 454–461, 2010.
- [190] S. Patra and L. Bruzzone, “A cluster-assumption based batch mode active learning technique,” *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1042–1048, 2012.
- [191] K. Pearson, “Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs,” *Proceedings of the royal society of London*, vol. 60, no. 359-367, pp. 489–498, 1896.
- [192] O. Pele and M. Werman, “The quadratic-chi histogram distance family,” in *European conference on computer vision*. Springer, 2010, pp. 749–762.
- [193] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 8, pp. 1226–1238, 2005.

- [194] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in 2009 IEEE International Conference on Multimedia and Expo. IEEE, 2009, pp. 1218–1221.
- [195] R. Petersen, "Use and misuse of multiple comparison procedures 1," *Agronomy Journal*, vol. 69, no. 2, pp. 205–208, 1977.
- [196] A. L. Pigot and R. S. Etienne, "A new dynamic null model for phylogenetic community structure," *Ecology letters*, vol. 18, no. 2, pp. 153–163, 2015.
- [197] J. Pinquier, J.-L. Rouas, and R. André-Obrecht, "Robust speech/music classification in audio documents," in *Seventh International Conference on Spoken Language Processing*, 2002.
- [198] R. C. Prim, "Shortest connection networks and some generalizations," *Bell system technical journal*, vol. 36, no. 6, pp. 1389–1401, 1957.
- [199] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine learning*, vol. 42, no. 3, pp. 203–231, 2001.
- [200] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 17–26.
- [201] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
- [202] L. R. Rabiner, R. W. Schafer et al., "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.
- [203] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.
- [204] C. Reimann, P. Filzmoser, K. Fabian, K. Hron, M. Birke, A. Demetriades, E. Dinelli, and A. Ladenberger, "The concept of compositional data analysis in practice—total major element concentrations in agricultural and grazing land soils of Europe." *The Science of the total environment*, vol. 426, pp. 196–210, Jun. 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0048969712002422>
- [205] M. Reindel and G. A. Wagner, *New Technologies for Archaeology: Multidisciplinary Investigations in Palpa and Nasca, Peru*. Springer, 2009.
- [206] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *International Journal of computers and communications*, vol. 5, no. 1, pp. 27–34, 2011.
- [207] ———, "Internal versus external cluster validation indexes," *International Journal of computers and communications*, vol. 5, no. 1, pp. 27–34, 2011.
- [208] R. Rennie, *The Facts on File Dictionary of Atomic and Nuclear Physics*, ser. *Facts on File science library*. Facts on File, 2003. [Online]. Available: <https://books.google.com.cy/books?id=zr5xQgAACAAJ>

- [209] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.
- [210] D. W. Roberts, "Ordination on the basis of fuzzy set theory," *Vegetatio*, vol. 66, no. 3, pp. 123–131, 1986.
- [211] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.
- [212] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *2006 IEEE Intelligent Transportation Systems Conference*. IEEE, 2006, pp. 733–738.
- [213] S. J. Russell, "Preliminary steps toward the automation of induction." in *AAAI*, 1986, pp. 477–484.
- [214] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.
- [215] S. J. Russell, *The use of knowledge in analogy and induction*. Morgan Kaufmann Publishers Inc., 1989.
- [216] D. Ryabko, "Pattern recognition for conditionally independent data," *The Journal of Machine Learning Research*, vol. 7, pp. 645–664, 2006.
- [217] S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data mining and knowledge discovery*, vol. 1, no. 3, pp. 317–328, 1997.
- [218] C. Sanden and J. Z. Zhang, "Enhancing multi-label music genre classification through ensemble techniques," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 705–714.
- [219] E. Savazzi and R. Reymont, *Aspects of Multivariate Statistical Analysis in Geology*. Elsevier, 1999.
- [220] L. L. Scharf, *Statistical signal processing*. Reading: Addison-Wesley, 1991, vol. Vol. 98.
- [221] E. D. Scheirer, "Sound scene segmentation by dynamic detection of correlogram comodulation," in the *International Joint Conference on AI Workshop on Computational Auditory Scene Analysis*, 1999.
- [222] S. Scherer, M. Glodek, G. Layher, M. Schels, M. Schmidt, T. Brosch, S. Tschechne, F. Schwenker, H. Neumann, and G. Palm, "A generic framework for the inference of user states in human computer interaction," *Journal on Multimodal User Interfaces*, vol. 6, no. 3-4, pp. 117–141, 2012.
- [223] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech & Language*, vol. 27, no. 1, pp. 263–287, 2013.

- [224] J. C. Schlimmer et al., “Efficiently inducing determinations: A complete and systematic search algorithm that uses optimal pruning,” in Proceedings of the 1993 International Conference on Machine Learning, 1993, pp. 284–290.
- [225] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [226] B. Schölkopf, A. J. Smola, F. Bach et al., *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- [227] F. Schwenker and E. Trentin, “Pattern classification and clustering: a review of partially supervised learning approaches,” *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014.
- [228] V. Sehwag, A. N. Bhagoji, L. Song, C. Sitawarin, D. Cullina, M. Chiang, and P. Mittal, “Analyzing the robustness of open-world machine learning,” in Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. ACM, 2019, pp. 105–116.
- [229] W. Siedlecki and J. Sklansky, “On automatic feature selection,” in *Handbook of Pattern Recognition and Computer Vision*. World Scientific, 1993, pp. 63–87.
- [230] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
- [231] J. A. Smith, J. E. Earis, and A. A. Woodcock, “Establishing a gold standard for manual cough counting: video versus digital audio recordings,” *Cough*, vol. 2, no. 1, p. 6, 2006.
- [232] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [233] L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo, “Supervised feature selection via dependence estimation,” in Proceedings of the 24th international conference on Machine learning. ACM, 2007, pp. 823–830.
- [234] Y. Song, F. Nie, and C. Zhang, “Semi-supervised sub-manifold discriminant analysis,” *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1806–1813, 2008.
- [235] C. R. Souza. (2010, March) Kernel functions for machine learning applications. [Online]. Available: <http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html>
- [236] J. C. Stegen, X. Lin, J. K. Fredrickson, X. Chen, D. W. Kennedy, C. J. Murray, M. L. Rockhold, and A. Konopka, “Quantifying community assembly processes and identifying features that impose them,” *The ISME journal*, vol. 7, no. 11, p. 2069, 2013.
- [237] R. Stiefelhagen, R. Bowers, and J. Fiscus, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. Springer, 2008, vol. 4625.

- [238] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, “Is robustness the cost of accuracy?—a comprehensive study on the robustness of 18 deep image classification models,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.
- [239] O. Sushkova, A. Morozov, A. Gabova, and A. Karabanov, “Investigation of the multiple comparisons problem in the analysis of the wave train electrical activity of muscles in parkinson’s disease patients,” in *Journal of Physics: Conference Series*, vol. 1368, no. 5. IOP Publishing, 2019, p. 052004.
- [240] P.-N. Tan, M. Steinbach, and V. Kumar, “Association analysis: basic concepts and algorithms,” *Introduction to data mining*, pp. 327–414, 2005.
- [241] J. Tang, S. Alelyani, and H. Liu, “Feature selection for classification: A review,” *Data classification: Algorithms and applications*, p. 37, 2014.
- [242] A. Temko, “Clear 2007 aed evaluation plan and workshop,” 2007.
- [243] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, “Acoustic event detection and classification in smart-room environments: Evaluation of chil project systems,” *Cough*, vol. 65, no. 48, p. 5, 2006.
- [244] A. Temko and C. Nadeu, “Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering,” in *Proceedings (ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–505.
- [245] A. Temko, C. Nadeu, and J.-I. Biel, “Acoustic event detection: Svm-based system and evaluation setup in clear’07,” in *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 354–363.
- [246] A. R. Templeton, “Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate bayesian computation,” *Molecular ecology*, vol. 18, no. 2, pp. 319–331, 2009.
- [247] F. A. Thabtah, P. Cowling, and Y. Peng, “Mmac: A new multi-class, multi-label associative classification approach,” in *Data Mining, 2004. ICDM’04. Fourth IEEE International Conference on*. IEEE, 2004, pp. 217–224.
- [248] B. Thakur and M. Mann, “Data mining with big data using c4. 5 and bayesian classifier,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 8, pp. 959–962, 2014.
- [249] S. Theodoridis and K. Koutroubas, “Feature generation ii,” *Pattern recognition*, vol. 2, pp. 269–320, 1999.
- [250] C. Thiel, S. Scherer, and F. Schwenker, “Fuzzy-input fuzzy-output one-against-all support vector machines,” in *Knowledge-based intelligent information and engineering systems*. Springer, 2007, pp. 156–165.
- [251] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, “Multi-label classification of music into emotions.” in *ISMIR*, vol. 8, 2008, pp. 325–330.

- [252] G. Tsoumakas, M. Zhang, and Z. Zhou, “Tutorial on learning from multi-label data [<http://www.ecmlpkdd2009.net/wp-content/uploads/2009/08/learningfrom-multi-label-data.pdf>],” in ECML/PKDD, 2009.
- [253] N. F. Udemba and O. Ibeneme, “Determination of human resource management practices of managers of small and medium enterprises in anambra state, nigeria,” *European Journal of Human Resource Management Studies*, 2020.
- [254] L. G. Underhill and M. Peisach, “Correspondence analysis and its application in multielemental trace analysis,” *Journal of trace and microprobe techniques* 3, pp. 41–65, 1985.
- [255] H. Vafaie and K. De Jong, “Genetic algorithms as a tool for feature selection in machine learning,” in *Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI’92*. IEEE, 1992, pp. 200–203.
- [256] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.
- [257] V. Vapnik, “Statistical learning theory. 1998,” 1998.
- [258] —, *The nature of statistical learning theory*. Springer Science & Business Media, 2013.
- [259] V. Vapnik and V. Vapnik, “Statistical learning theory,” 1998.
- [260] E. Veklerov and J. Llacer, “Stopping rule for the mle algorithm based on statistical hypothesis testing,” *IEEE Transactions on Medical Imaging*, vol. 6, no. 4, pp. 313–319, 1987.
- [261] T. Villmann and H.-U. Bauer, “Applications of the growing self-organizing map,” *Neurocomputing*, vol. 21, no. 1-3, pp. 91–100, 1998.
- [262] P. Viola and M. Jones, “Fast and robust classification using asymmetric adaboost and a detector cascade,” in *Advances in neural information processing systems*, 2002, pp. 1311–1318.
- [263] J. Wang and I. C. Paschalidis, “Statistical traffic anomaly detection in time-varying communication networks,” *IEEE Transactions on Control of Network Systems*, vol. 2, no. 2, pp. 100–111, 2014.
- [264] M. Wang, X. Zhou, and T.-S. Chua, “Automatic image annotation via local multi-label classification,” in *Proceedings of the 2008 international conference on Content-based image and video retrieval*. ACM, 2008, pp. 17–26.
- [265] L. Wasserman, “The role of assumptions in machine learning and statistics: Don’t drink the koolaid!” 2015.
- [266] C. O. Webb, D. D. Ackerly, M. A. McPeck, and M. J. Donoghue, “Phylogenies and community ecology,” *Annual review of ecology and systematics*, vol. 33, no. 1, pp. 475–505, 2002.

- [267] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, “Use of the zero-norm with linear models and kernel methods,” *Journal of machine learning research*, vol. 3, no. Mar, pp. 1439–1461, 2003.
- [268] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
- [269] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
- [270] A. Woronow and K. M. Love, “Quantifying and testing differences among means of compositional data suites,” *Mathematical Geology*, vol. 22, no. 7, pp. 837–852, 1990.
- [271] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip et al., “Top 10 algorithms in data mining,” *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.
- [272] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, “Distance metric learning with application to clustering with side-information,” *Advances in neural information processing systems*, vol. 15, pp. 505–512, 2003.
- [273] L. Xu, P. Yan, and T. Chang, “Best first strategy for feature selection,” in [1988 Proceedings] 9th International Conference on Pattern Recognition. IEEE, 1988, pp. 706–708.
- [274] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, “Audio keywords generation for sports video analysis,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 4, no. 2, p. 11, 2008.
- [275] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, “Discriminative semi-supervised feature selection via manifold regularization,” *IEEE Transactions on Neural networks*, vol. 21, no. 7, pp. 1033–1047, 2010.
- [276] L. Yang and R. Jin, “Distance metric learning: A comprehensive survey,” *Michigan State University*, vol. 2, 2006.
- [277] H. Zhang, “The optimality of naive bayes,” *AA*, vol. 1, no. 2, p. 3, 2004.
- [278] M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 8, pp. 1819–1837, 2014.
- [279] T. Zhang and C.-C. J. Kuo, “Hierarchical system for content-based audio classification and retrieval,” in *Multimedia Storage and Archiving Systems III*, vol. 3527. International Society for Optics and Photonics, 1998, pp. 398–409.
- [280] Z. Zhao and H. Liu, “Semi-supervised feature selection via spectral analysis,” in *Proceedings of the 2007 SIAM international conference on data mining*. SIAM, 2007, pp. 641–646.
- [281] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, “Hmm-based acoustic event detection with adaboost feature selection,” in *Multimodal technologies for perception of humans*. Springer, 2007, pp. 345–353.

- [282] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.
- [283] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 17–20.
- [284] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine." *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, 1993.