MACHINE LEARNING FOR PRIVACY PRESERVING DATA PUBLISHING AND THE ANALYSIS OF CATEGORICAL DATA IN THE MEDICAL DOMAIN

Aristos Aristodimou

University of Cyprus, 2019

Με την εισαγωγή της επιστήμης της πληροφορικής στον τομέα της υγείας, πολλές οντότητες που συσχετίζονται με την υγεία, έχουν στην κατοχή τους τεράστιο όγκο από δεδομένα ασθενών. Παρά το γεγονός ότι η κοινοποίηση αυτών των δεδομένων σε ερευνητές μπορεί να αυξήσει την πιθανότητα ανακάλυψης καινοτόμων ευρημάτων, αυτό δεν είναι δυνατό λόγο νομικών και ηθικών ζητημάτων. Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί σε αυτά τα δεδομένα για τον εντοπισμό των παραγόντων που αυξάνουν ή μειώνουν το ρίσκο κάποιος να αποκτήσει μια ασθένεια, αλλά οποιοσδήποτε αλγόριθμος που θα αναλύσει αυτά τα δεδομένα θα πρέπει να λαμβάνει υπόψη ότι οι περισσότερες κοινές ασθένειες επηρεάζονται από πολλαπλές γονιδιακές αλληλεπιδράσεις και αλληλεπιδράσεις με το περιβάλλον. Επομένως, θα πρέπει να χρησιμοποιηθούν αλγόριθμοι που επιτρέπουν την εύρεση τέτοιων αλληλεπιδράσεων. Στη διατριβή, αρχικά παρουσιάζεται ένας νέος αλγόριθμος για την ανωνυμοποίηση δεδομένων τα οποία έχουν διακριτές τιμές, μέσω κ-ανωνυμίας και αλγόριθμου επιλογής μεταβλητών, για προβλήματα ταξινόμησης. Ο αλγόριθμος αξιολογήθηκε σε διάφορα είδη ιατρικών δεδομένων και στην πλειονότητα των πειραμάτων τα ανώνυμα δεδομένα που παράχθηκαν, είχαν παρόμοια ή μεγαλύτερη ακρίβεια στην ταξινόμηση σε σχέση με τη χρήση των αρχικών μη-ανώνυμων δεδομένων. Στη συνέχεια παρουσιάζεται ένας καινούριος αλγόριθμος για τη μετατροπή συνεχών μεταβλητών σε διακριτές, με βάση την πυκνότητα των τιμών των μεταβλητών. Η μέθοδος αυτή έχει παρόμοιες επιδόσεις με τους αλγόριθμους που είναι ευρέως χρησιμοποιημένοι στον τομέα, και έχει το πλεονέκτημα ότι είναι υπολογιστικά αποδοτικός και μπορεί να χρησιμοποιηθεί σε μεγάλα δεδομένα. Για την αναγνώριση προτύπων σημειακών νουκλεοτιδικών πολυμορφισμών που συσχετίζονται με εμφάνιση ή όχι μιας ασθένειας, παρουσιάζεται ένας Χάρτης Αυτο-οργάνωσης για διακριτά δεδομένα. Η μέθοδος αυτή εφαρμόστηκε σε γενετικά δεδομένα και η κατηγοριοποίηση των δεδομένων που δημιούργησε ήταν στατιστικά σημαντική και αποκάλυψε ενδιαφέροντα πρότυπα που ήταν διαφορετικά μεταξύ των κατηγοριών που αντιπροσώπευαν ασθενείς και υγιή άτομα. Επίσης, προτείνεται ένα πλαίσιο για την αποτελεσματική ανακάλυψη ν-αλληλεπιδράσεων. Το πλαίσιο αυτό χρησιμοποιεί αλγόριθμους μηχανικής μάθησης για την μείωση του αριθμού των μεταβλητών ενός προβλήματος και για τη μείωση της διάστασης των μεταβλητών μετατρέποντάς τις σε δυαδική μορφή. Αυτό επιτρέπει την μείωση του προβλήματος των πολλαπλών συγκρίσεων και επίσης μειώνει τους βαθμούς ελευθερίας των στατιστικών υποθέσεων, το οποίο αυξάνει την στατιστική δύναμη για την αναγνώριση των ν-αλληλεπιδράσεων που συσχετίζονται με εμφάνιση ή όχι μιας ασθένειας. Τα αποτελέσματα δείχνουν ότι με τη νέα κωδικοποίηση, το προτεινόμενο πλαίσιο ήταν σε θέση να αναγνωρίσει περισσότερες στατιστικά σημαντικές ν-αλληλεπιδράσεις σε σύγκριση με τη χρήση της αρχικής κωδικοποίησης των μεταβλητών.

Aristos Aristodimou––University of Cyprus, 2019

In recent years, with the infiltration of information technology in healthcare, many healthcare related entities, have vast amounts of patients' data. Although sharing such data can increase the likelihood of identifying novel findings or even replicating existing research results, this is not happening due to legal and ethical issues. Machine learning can be used on such datasets to identify risk factors that can be used to improve our lives, but any algorithms that will analyze such data should take into consideration that most common diseases are influenced by multiple gene interactions and interactions with the environment. Hence they should use models that allow the finding of such multivariate associations. This thesis initially presents a novel algorithm for anonymizing categorical data with k-anonymity and performing feature selection for classification tasks. The algorithm was evaluated on various medical datasets and in the majority of the evaluated test cases the produced anonymized data had similar or better accuracies than using the full datasets. Additionally, a novel density based discretization algorithm is presented that has similar performance with state of the art algorithms while being computationally efficient and suitable for big data. For pattern recognition of n-SNP associations in case/control data, a Self Organizing Map for nominal categorical data is presented, which was able to produce statistically significant clustering revealing some interesting patterns between the clusters of cases and controls. Finally, a framework for efficient n-Way interaction testing is presented that uses machine learning to reduce the dimensionality of the data and to produce a targeted binary encoding of the features. This enables the reduction of the multiple testing problem and the degrees of freedom of the statistical tests applied for interaction testing, and hence increases the statistical power of the performed analysis. Results indicate that with the new encoding, the proposed framework was able to identify more statistically significant interactions compared to using the initial encoding of the features.

MACHINE LEARNING FOR PRIVACY PRESERVING DATA PUBLISHING AND THE ANALYSIS OF CATEGORICAL DATA IN THE MEDICAL DOMAIN

Aristos Aristodimou

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Cyprus

Recommended for Acceptance

by the Department of Computer Science

November, 2019

APPROVAL PAGE

Doctor of Philosophy Dissertation

MACHINE LEARNING FOR PRIVACY PRESERVING DATA PUBLISHING AND THE ANALYSIS OF CATEGORICAL DATA IN THE MEDICAL DOMAIN

Presented by

Aristos Aristodimou

Research Supervisor ______________________
Constantinos S. Pattichis

Committee Member ______________________
Chris Christodoulou

Committee Member ______________________
Christos N. Schizas

Committee Member ______________________
George Spyrou

Committee Member ______________________
Dimitrios Koutsouris

University of Cyprus

November, 2019

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to my Ph.D. supervisor, Dr. Constantinos S. Pattichis, for his continuous support and valuable feedback during these years. He has always been there for me, offering me his valuable knowledge, which helped me become a better researcher and made this thesis dissertation possible.

I would also like to thank Dr. Athos Antoniades, who gave me the chance to get involved in his research project when I was still an undergraduate and introduced me to the world of genetics, interaction testing and of-course his favorite programming language C++. He has helped me as a mentor, co-worker, and friend to become a better professional and person.

I also thank my committee members, Dr. Chris Christodoulou, Dr. Christos N. Schizas, Dr. George Spyrou and Dr. Dimitrios Koutsouris for their insightful comments and feedback, which incentivized me to widen my research from various perspectives.

Last but not least, I would like to thank my family: my grandparents Dimitris and Stravroulla, my parents Pantelis and Eleni, my sister Dimitra, my wife Anna and my son Alexandros for all of their love and support throughout these years.

CREDITS

Part of the work was conducted as part of the project *A next-generation secure linked data medical information space for semantically-interconnecting electronic health records and clinical trials systems advancing patients safety in clinical research (Linked2Safety)* that received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No 288328.

Part of the work was co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research Promotion Foundation (Project Cy-Tera ΝΕΑ ΥΠΟΔΟΜΗ/ΣΤΡΑΤΗ/0308/31).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

1-NN 1-Nearest Neighbor

AE Adverse Event

ANOVA Analysis of Variance

BMU Best Matching Unit

CAIM Class-Attribute Interdependence Maximization

CI Confidence Interval

CNN Convolutional Neural Network

CV Cross Validation

DBAD Density Based Discretization

DMPD Data Mining Privacy by Decomposition

EDC Electronic Data Capture

EHR Electronic Health Record

FN False Negative

FP False Positive

FSFS Forward Sequential Feature Selection

GENN Grammatical Evolution Neural Network

GPNN Genetic Programming Neural Network

GWAS Genome Wide Association Study

HLA Human Leukocyte Antigen

HWE Hardy Weinberg Equilibrium

kACTUS k-Anonymity Classification Tree-based Suppression

kPB-MS k-anonymity through Pattern Based Multidimensional Suppression

MAF Minor Allele Frequency

MDLP Minimum Description Length Principle

MDR Multifactor Dimensionality Reduction

MedDRA Medical Dictionary for Regulatory Activities

MPI Message Passing Interface

MS Multiple Sclerosis

NB Naive Bayes

NCSOM Numerical and Categorical Self Organizing Map

NN Neural Network

NPA Net Percent Agreement

PB-FSS Pattern Based Feature Subset Selection

PBCA Pattern Based Classification Accuracy

PPDP Privacy Preserving Data Publishing

RDF Resource Description Framework

RF Random Forest

RFE-SVM Recursive Feature Elimination Support Vector Machine

RJ Random Jungle

SD Standard Deviation

SNP Single Nucleotide Polymorphism

SNPInterForest Single Nucleotide Polymorphism Interaction Forest

SOM Self Organizing Map

SPECT Single Proton Emission Computed Tomography

SVM Support Vector Machine

TDR Top-Down Refinement

TN True Negative

TP True Positive

LIST OF SYMBOLS

x	single value variable
$\mathbf{x}$	vector
$\mathbf{x}^y$	the elements of $\mathbf{x}$ that belong to class y
$\lvert\mathbf{x}\rvert$	number of elements in vector $\mathbf{x}$
$\boldsymbol{x}$	vector of objects
$\boldsymbol{x}_i$	the i_{th} object of $\boldsymbol{x}$
$\boldsymbol{x}_i^n$	the size of the i_{th} object of $\boldsymbol{x}$
$\boldsymbol{x}_i^{left}$	the left interval of the i_{th} object of x
$\boldsymbol{x}_i^{right}$	the right interval of the i_{th} object of $\boldsymbol{x}$
X	matrix
$X_{i,j}$	the j_{th} feature of the i_{th} instance of matrix X

Chapter 1

Introduction

1.1 Problem Statement

In recent years, with the infiltration of information technology in healthcare, many healthcare related entities such as hospitals and pharmaceutical companies, have vast amounts of patients' data. Such datasets may contain various types of data, such as environmental, phenotypic, imaging and -omics data, which can provide useful information for identifying factors using machine learning that can improve our lives. For example, imaging data could be used to predict if a person is on early stages of dementia or the risk of developing cancer, whereas environmental and phenotypic data could provide risk factors for a person to develop a cardiovascular disease and hence have the ability to take precautions for reducing the risks and having a better and longer life. Genetic data can change the whole scene in the current medical industry, since they can make personalized medicine a reality.

Although it is clear that sharing such data can increase the likelihood of identifying novel findings or even replicating existing research results, this is not happening due to legal and ethical issues. Attempts have been made by research programs like Linked2Safety [2, 1], for merging data across different healthcare entities and preserve individuals privacy, but one needs to also decide which data to share, so that the likelihood that the shared data will lead to new findings is increased. The problem is more evident in the case of genetic data, since they can reveal the identity of an individual and possible risk factors that s/he may have. Hence they are highly sensitive data and caution is needed even if a part of them is shared.[3]

Machine learning can be used for the analysis of such data, but such algorithms should take into consideration that most common diseases are influenced by multiple gene interactions and interactions with the environment [5]. Furthermore, univariate analyses in such diseases may not replicate their results across multiple samples, due to the effect of epistasis and other phenomena [6]. Hence, any algorithm that will be used as a preprocessing and processing step for the analysis of this data, should allow the finding of such multivariate associations.

Additionally, due to the high dimensionality of these datasets, the algorithms that will process them should be efficient and be able to harness the computational power of high performance computing. This means that the algorithms need to be as computationally efficient as possible, but additionally be easily parallelizable so as to be executed on multiple nodes and cores.

1.2 Contribution

This thesis focuses on the sharing and the analysis of categorical data in the medical domain and contributes to various steps that could be needed in the analysis process of such data. To be able to include continuous data as well in the analysis process, a novel discretization algorithm for converting continuous data to categorical is also proposed. Assuming that all data are categorical or have been converted to categorical, the proposed algorithms in this thesis and frameworks can be applied to them. Then this thesis contributes to privacy preserving data publishing, by providing a novel algorithm for the anonymization of data to be able to share such data with other researchers and to be able to combine data from different data providers. This can increase the statistical power of the analysis and the likelihood of identifying novel findings or replicating existing results. If needed then, before analyzing a dataset, one can discretize and anonymize the data at hand using the proposed algorithms. The next two areas in which this thesis contributes are in the analysis of data and specifically in the identification of n-SNP patterns and n-Way interactions in case/control studies. In both cases, the algorithms and framework proposed, take into consideration the use of methods that allow the identification of interacting factors since, in most common complex diseases, such interacting factors are influenced by them. Additionally, all of the proposed algorithms are capable to handle big data since they are either available in parallel implementations or are computationally efficient. In the next subsections, a summary of the contribution in each area is provided.

1.2.1 Privacy Preserving Data Publishing

The Linked2Safety project [2, 1] provides a framework for connecting and interlinking data from different sources while preserving data privacy and conforming to all legal and ethical issues relevant to each dataset. The privacy preserving methodology is based on the creation of data cubes and applying cell suppression and perturbation on them for removing patients data that can be reidentified. This is applied on categorical data so that the data cubes can be created. Additionally the data cubes are semantically enriched with the use of medical taxonomies and ontologies, to allow the interlinking of data from different sites. The data can then be analyzed in the data analysis space, using statistical and data mining methods.

One problem identified in the Linked2Safety project, was the selection of the variables that would be included in the data cubes for the analysis. For this reason a novel algorithm for anonymizing categorical data with k-anonymity and performing feature selection for classification tasks was proposed [3]. It is based on a novel fast pattern based evaluation measure that can be computed in O(n) and uses a greedy forward selection for selecting the best features, which can also be parallelized. The algorithm also tries to reduce the probability of removing data that can affect the results of future analysis using statistical measures. The algorithm was evaluated on various medical datasets with varying complexity, using four different classifiers. In 74% of the evaluated cases the produced anonymised data, had similar or better accuracies than using the full datasets.

1.2.2 Data Discretization

Another problem identified in the Linked2Safety project was the need to manually convert the continuous data to categorical, based on the expertise of the data owners. This approach required a good communication among the different sites, so as to use the same intervals in the variables that would be shared. Discretization is a preprocessing technique used for converting continuous features into categorical. This step is essential for processing algorithms that cannot handle continuous data and can also improve the performance and the interpretation of results of algorithms that can directly use continuous data as input. Moreover, the Self Organizing Map algorithm and the Privacy Preserving Data Publishing algorithm require categorical data and hence it was of interest to find an effective and efficient algorithm for converting continuous data to categorical. To address this issue, a novel density based discretization (DBAD) algorithm was proposed. It combines the simple equal width binning approach with a statistical test for identifying the intervals in which the density of the data of each class changes for producing the final bins. The algorithm has similar performance with state of the art algorithms, but has the advantage of being computationally efficient since it can discretize a variable in O(n) time and it can be easily parallelized. The parallel version of DBAD, shows almost linear speedup for an MPI implementation (9.6x for 10 nodes), while a hybrid MPI/OpenMP implementation improved execution time by 35.3x for

10 nodes and 6 threads per node. The algorithm can be applied to datasets that will be used in classification tasks and unsupervised tasks and can also be used for histogram visualization.

1.2.3 Pattern Recognition of n-SNP Associations

As mentioned earlier, common diseases can be affected by multiple interactions and hence algorithms that can help identify them should be used for analysing them. For this reason, a Self Organizing Map (SOM) that can handle nominal categorical data [4] was used for the first time on clustering Single Nucleotide Polymorphisms (SNPs) of subjects with and without Multiple Sclerosis. It was also proposed to use the χ^2 statistic for selecting the map size of the network instead of the traditional measures. The produced clusters were statistically significant and revealed interesting patterns between cases and controls, indicating that this method could prove of significant value in future multi-loci association testing. Additionally, a parallel version of the algorithm is available that can be used for the analysis of big data.

1.2.4 A Framework for Efficient n-Way Interaction Testing

As it has already been mentioned, complex interactions can affect common diseases, but with the high dimensionality of genetic data and usually the small sample size, an efficient n-Way interaction testing approach is needed. A framework that tackles this problem was proposed that comprises of four steps. The first step carries out quality control analysis so that the data used are not erroneous. The second step uses feature selection to remove any redundant features. By reducing the dimensionality of the data, the risk of overfitting in the next steps is reduced and it is easier for machine learning algorithms to identify the true signals from noise. In addition less computational power is required by the algorithms and the multiple testing problem is reduced. In the third step, clustering is used for grouping subjects with similar genetic patterns together in an attempt to identify clusters that have a majority of cases and controls and that have distinctive patterns that can separate the two classes. In the final step, the genotypes are converted to binary variables, based on the patterns of the clusters before testing for n-Way interactions. This reduces the degrees of freedom of the statistical tests applied for interaction testing and hence increases the statistical power. Results indicate that with the new encoding, the proposed framework was able to identify more statistically significant interactions compared to using the initial encoding of the genotypes.

1.2.5 Publications

The research work of this proposal, has produced 2 Journal publications and 8 conference papers, which are listed in Appendix A. Additionally 2 more Journal publications are under review.

1.3 Structure of this Proposal

This thesis is comprised of 8 chapters. Chapter 2 introduces the Linked2Safety project, which is focused on the anonymization and semantic interconnection of medical information across different entities for advancing patients' safety through its innovative platform. Chapter 3, introduces a novel privacy preserving data publishing algorithm that uses k-anonymity and feature selection and that can automate the selection of the important features that will be anonymized and published for analysis. Chapter 4, presents a novel fast discretization algorithm, which creates its bins based on the density of the data using a statistical approach. The next chapter is on a methodology for identifying n-SNP associations between cases and controls using a Self Organizing Map for categorical data, whereas the 7th chapter presents a framework for the efficient identification of n-Way interactions. Finally, in the last chapter, the concluding remarks and the proposed future work are presented.

Chapter 2

Linked2Safety: A secure linked data medical information space for semantically-interconnecting EHRs advancing patients' safety in medical research [1, 2]

2.1 Introduction

In this chapter Linked2Safety [1, 2] is presented, a project funded under the FP7 scheme of the European Commission focused on some key challenges and potential solutions that have been identified through the work performed in the project. The vision of the project was to advance clinical practice and accelerate medical research, by providing pharmaceutical companies, healthcare professionals and patients with an innovative semantic interoperability framework facilitating the efficient and homogenized access to distributed Electronic Health Records (EHRs).

EHRs contain an increasing wealth of medical information and they have the potential to contribute significantly and advance medical research, as well as improve health policies, providing society with additional benefits [7, 8, 9, 10]. However, the European healthcare information space is fragmented due to the lack of legal and technical standards, cost effective platforms, and sustainable business models. The key factors that define the work done in this and any other project that attempts to merge or share medical data between multiple sources are legal and ethical issues, patient privacy protection and statistical power.

Linked2Safety, developed a platform for analysing EHRs from multiple distributed institutions, while strictly adhering to the legal and ethical requirements as defined by each data provider at EU level. The primarily objective of analysing EHRs from multiple institutions, arises from the need to increase the total number of subjects that are included in each analyses, thus increasing the statistical power of detecting true positive effects. This may in turn enable the identification of key biomarkers, which may lead to new drugs, the identification of adverse events that may be associated with a specific drug or family of drugs, and the reduction of costs associated with the setting up of clinical trials [11, 12, 13, 14]. The costs of clinical trials can be reduced, since it will make the process of identifying sites and the number of subjects to be included in a clinical trial from each site easier.

The structure of the chapter is as follows. Section 2.2 covers the legal and ethical issues of the project, whereas Section 2.3 and Section 2.4 present the Linked2Safety platform and architecture respectively. Section 2.5 presents the three showcases used for demonstrating the use of the platform. Finally in Section2.9 the concluding remarks are provided.

2.2 Legal and Ethical issues

Within the Linked2Safety project, patients' personal health are processed, which prerequisites the study of the corresponding national and European legislative framework, and the establishment of specific legal and ethical requirements for the platform that will be developed and operating on these data.

According to Art.17 para. 1 of the European Data Protection Directive [15], the data controller as well as the data processor must implement appropriate technical and organizational measures

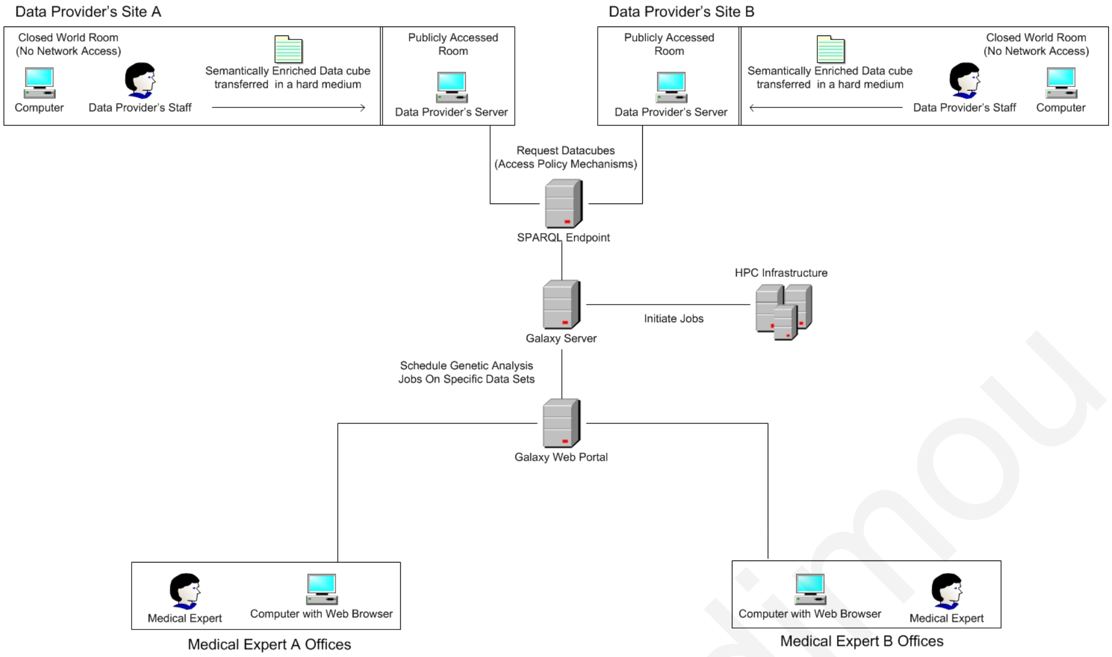


Figure 1: The Linked2Safety platform. The data are initially anonymised and semantically enriched before making them available for access by a SPARQL endpoint. The medical expert connects to a Galaxy web portal, which schedules any analysis that the user wants to perform. The analysis is then performed on an HPC infrastructure and the results are returned to the user through the Galaxy web portal.

to protect the data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access and against all other unlawful forms of processing. At the same time, the data subject has several rights that must be respected, such as the right to be informed, the right of access, the right of rectification, erasure or blocking and the right to object.

From the ethical perspective, the processing of health data at EU level requires the consent of the data subject [16], though in certain cases and under specific conditions this might not be needed. For instance, the national laws of the clinical pilot partners of Linked2Safety in Greece [17], Cyprus [18] and Switzerland [19] declare that the use and processing of health data without the consent of the data subject is possible, if and only if, it is for research purposes, the data are properly anonymised, and analysis is done in aggregated level.

2.3 Linked2Safety Platform

The Linked2Safety architecture addresses legal and ethical issues while enabling the analysis of electronic health data across multiple data sites with different owners and with potentially diverse legal and ethical requirements.

With those two facets of the problem, a novel idea for addressing both legal and ethical problems (at a commercial and not just a research level), as well as the need to merge the data from multiple sites simultaneously is introduced. All processing and analysis on medical data is done on aggregated data from each site. This is nothing new, since aggregated results from the analysis of electronic health records are published in the majority of medical research publications that involve medical studies. In Linked2Safety however the system enables future analysis of the aggregated data using methodologies or ways of analysis that may have not been considered at the time of aggregating the data. The proposed method that enables this is the use of data cubes [20]. This method aggregates values across many dimensions together, resulting in multidimensional data cubes. Getting a contingency table from any subset of the variables within the data cube can easily be achieved by adding the values across the excluded dimensions.

For tackling the issue of accessing electronic health records that contain potentially identifiable personal medical data, the concept of a ”closed-world” room was introduced (a room located within a data provider’s premises, featuring the required hardware infrastructure to process EHR’s isolated from any kind of network connections).

The physical access to this machinery within the room is allowed only to specific personnel of the corresponding data provider and it is off line to the outside world. The data provider’s staff executes a program on the computers located in the closed room, that aggregates the data generating the data cubes. The program offers the option to the data provider to limit the way that the data are aggregated so that any legal and ethical issues that may relate to the type of analyses that may be performed on the data can be addressed. Quality control is also performed on the data by the program, so that the likelihood of reverse engineering the data of a single subject or a group of subjects is minimized. As illustrated in Figure 1, only the aggregated data (data cubes) are physically carried outside the closed-world room to a server that is accessible by the rest of the Linked2Safety infrastructure. The data cubes are in Resource Description Framework (RDF) format and are accessed through a SPARQL endpoint, whereas all of the tools for analysing the aggregated data are on a dedicated Galaxy server [21, 22]. The analysis of the data is available only to Linked2Safety users through a Galaxy web portal.

The operational procedures for the creation and semantic enrichment of the data cubes are as follows:

- The data provider's staff after reviewing the legal and ethical requirements for their data, make a decision on what data to include in Linked2Safety and what parameters they need to define for the creation of the aggregated data.

- A member of the staff of the data provider enters the ”closed-world” room, where the data are maintained and performs the aggregation of the data, which creates the data cubes. This step includes the quality assurance and filtering of the data, based on the predefined settings of the previous step.

- The produced data cubes are stored in RDF format and are verified that they do not contain any personal medical records.

- The final data cubes are transferred to a server that is accessible by the Linked2Safety platform, outside the ”closed-world” room.

- The data cubes are semantically enriched and made available to the rest of the Linked2Safety platform.

The above steps are carried out by all sites that maintain clinical data, so as to include in the Linked2Safety platform aggregated data from multiple independent studies. In cases where the definitions of the variables overlap among sites, they are recorded in a way that enables them to be analysed and merged so as to increase the statistical power of detecting true positive results. These data are then available to researchers with access to the Linked2Safety platform for analysing them.

2.4 Linked2Safety Architecture

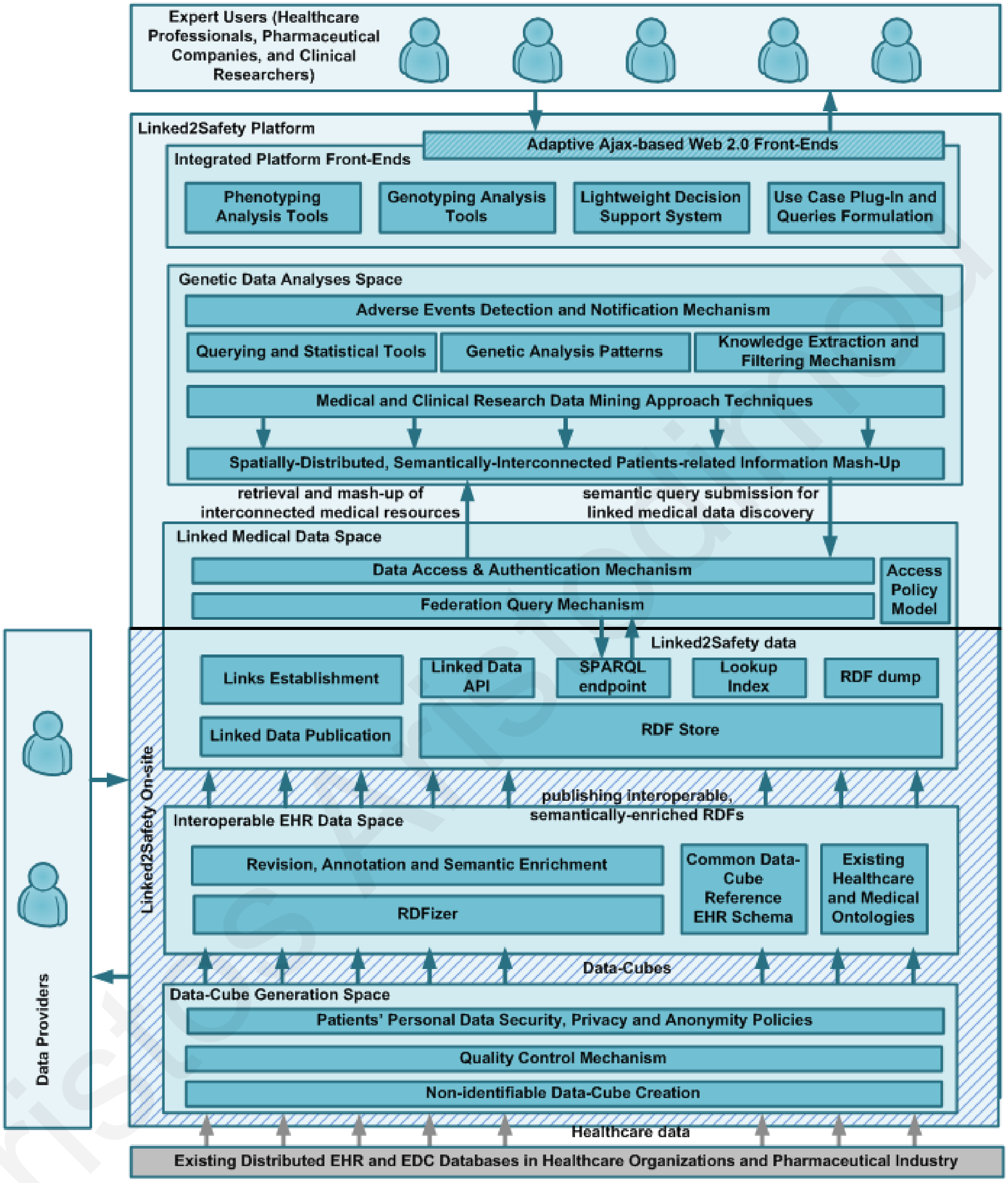

Expert Users (Healthcare Professionals, Pharmaceutical Companies, and Clinical Researchers)
Linked2Safety Platform
Adaptive Ajax-based Web 2.0 Front-Ends
Integrated Platform Front-Ends
Phenotyping Analysis Tools
Genotyping Analysis Tools
Lightweight Decision Support System
Use Case Plug-In and Queries Formulation
Genetic Data Analyses Space
Adverse Events Detection and Notification Mechanism
Querying and Statistical Tools
Genetic Analysis Patterns
Knowledge Extraction and Filtering Mechanism
Medical and Clinical Research Data Mining Approach Techniques
Spatially-Distributed, Semantically-Interconnected Patients-related Information Mash-Up
retrieval and mash-up of interconnected medical resources
semantic query submission for linked medical data discovery
Linked Medical Data Space
Data Access & Authentication Mechanism
Federation Query Mechanism
Access Policy Model
Linked2Safety data
Links Establishment
Linked Data API
SPARQL endpoint
Lookup Index
RDF dump
Linked Data Publication
RDF Store
Linked2Safety On-site
publishing interoperable, semantically-enriched RDFs
Interoperable EHR Data Space
Revision, Annotation and Semantic Enrichment
RDFizer
Common Data-Cube Reference EHR Schema
Existing Healthcare and Medical Ontologies
Data-Cubes
Data-Cube Generation Space
Patients' Personal Data Security, Privacy and Anonymity Policies
Quality Control Mechanism
Non-identifiable Data-Cube Creation
Healthcare data
Existing Distributed EHR and EDC Databases in Healthcare Organizations and Pharmaceutical Industry
Data Providers

Figure 2: The Linked2Safety architecture consists of four spaces. The *data cube generation space* creates the data cubes for anonymising the patients personal data. The *interoperable EHR data space* provides the tools for transforming the data cube information to a common referenced data cube format. The *linked medical data space* provides a secure knowledge base of semantically interconnected data cube related information resources. The fourth space, *genetic analysis space*, provides the tools for data mining and statistical analysis of the medical data.

Figure 2 portraits the Linked2Safety's architecture. As can be seen, the platform consists of four main spaces, which are described below.

2.4.1 Data Cube Generation Space

The *data cube generation space* provides the tools for transforming the proprietary EHR and Electronic Data Capture (EDC) data to the data cube structure. Specifically, the locally stored EHR and EDC data are aligned, mapped and transformed to a reference EHR schema. Then, this commonly referred EHR data are used to create the data cubes.

2.4.2 Interoperable EHR Data Space

The *interoperable EHR data space* provides a tool set to make the transition of data cubes from the "closed-world" of each clinical data provider to an open data environment accessible to all partners based on policies that enforce strong data security, privacy and anonymity. Thus, its responsibility is to transform the data cube information to a common referenced data cube format by means of a semantic EHR model, named common data cube reference EHR ontology. Moreover, the interoperable EHR data space provides the mechanisms for the semantic enrichment of the standardised data cubes with the use of appropriate, globally available healthcare and medical taxonomies and ontologies, enabling the delivery of machine interpretable information regarding their structure and content.

2.4.3 Linked Medical Data Space

The *linked medical data space* implements a secure knowledge base of semantically interconnected data cube related information resources. It also provides the mechanisms and tools required for publishing and interlinking the common referenced data cubes from different medical data providers. Access to this data is governed by adaptable access policies and mechanisms. In this way, the clinical research community is going to have homogenised access to the available anonymised patient related information needed, to perform complex data mining operations.

2.4.4 Genetic Data Analysis Space

The *genetic data analysis space* provides a scalable infrastructure for medical data mining, empowered by a set of algorithms and models. These methods are applied in the semantically-interlinked data cubes containing anonymised patient's health records in order to analyse the associations among the genetic, environmental and phenotypical data related to identified and reported Adverse Events (AEs). Thus, clinical researchers and healthcare professionals are provided with an advanced genetic analysis statistical and data mining toolset, focusing on advancing patients' safety through the analysis of bio-markers associated to identified AEs and the proactive exclusion of specific patients' profiles from the wide patients' selection process.

My main work in the Linked2Safety project was in this space, in which I was also the leader. Part of my work was the design of the architecture of the space, the selection and implementation of the algorithms that would be included in the analysis space and the selection of the platform that would integrate all of the algorithms. This space was implemented using the Galaxy workflow management system and it is comprised of five steps, which are illustrated in Figure 3. Galaxy allows the creation of different workflows that can be stored and used with new data. Additionally, it allows the sharing of such workflows with other researchers, which can promote the collaboration and transfer of analysis workflows between different research groups. The input data are retrieved using a SPARQL query engine that accesses the anonymized RDF data cubes. For pre-processing, multiple quality control methods are available such as handling missing data and performing the Hardy Weinberg Equilibrium test. Additionally, feature selection algorithms are also available for the pre-processing step based on rough set theory and statistical measures such as the χ^2 test and information gain. This helps to reduce the feature space of the analysis and helps focus on the most informative features. Additionally, this allows the creation of simpler models that can be interpreted for drug discovery and the creation of human-readable association rules. In the processing step, data mining algorithms and statistical methods are provided for analyzing the data. These include Random Forests [23], C4.5 [24], logistic regression and statistical tests that can be used to test for associations between the predictors and the target variable. In the post-processing step, any statistically significant identified rules are automatically stored in the knowledge base of the platform, which can be used to alert for any possible adverse effects in specific subjects. Finally, in the last step, the results of the performed analyses are output in various forms such as plots and texts.

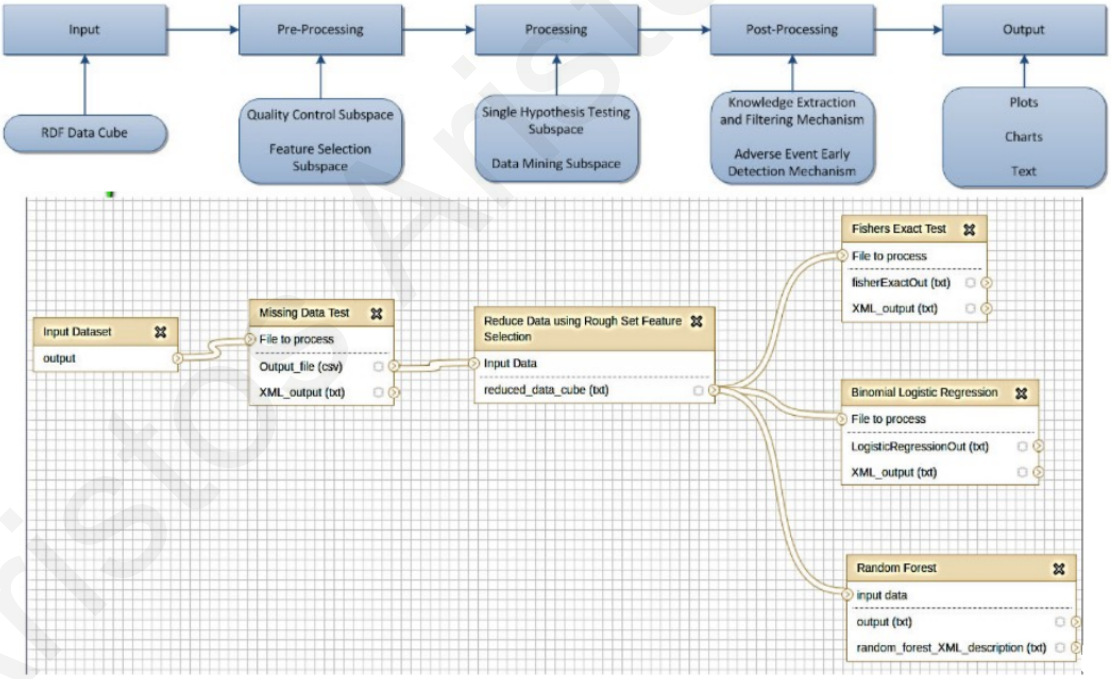


Figure 3: The Linked2Safety's Genetics Data Analysis Space workflow steps and an example of a workflow in Galaxy.

2.5 Usage Scenarios

Although the Linked2Safety platform will enable many diverse types of analysis that may help in multiple facets of research in medicine, three were identified and acted as showcases to demonstrate the use of the platform when it was completed.

2.5.1 Subject Selection for a Phase III Clinical Trial

Recruitment of investigators, clinical sites, and study participants is a constant concern during the course of any clinical trial. Recruiting participants for clinical trials is a time-consuming, intensive process which constitutes the critical first step for a study's ultimate success. Once investigators and clinical sites are selected, participants recruitment and retention is performed. Clinical researchers then determine whether clinical studies adhere to time-lines. Depending upon the size of the trial, the condition or disease being investigated, and the participant population available, subject recruitment often proves to be a time-consuming step for many clinical trials.

Clinical studies are no longer exclusively centralised in university and research hospitals. Many clinical trials have been relocated to private practice or small clinical research centres, often spread across the continent in pan-European trials. Investigators and academic research organisations are also implementing global clinical trials and mining databases in search of physicians who treat patients eligible for trials. Sponsors and clinical research organisations usually rely on a network of known investigators and sites, tending to go to the sites that they have been working with in the past. To succeed in the identification of superior sites in terms of quality and subject recruitment, they maintain active investigator databases for Europe and the United States. Information typically included in such databases is: contact details, area of research interest, investigator's experience with clinical trials (phase of trials, number of patients enrolled at site per trials, classes of drug/devices tested), site infrastructure and capacity (equipment, internal logistics, patient database), etc.

Moreover, another way to identify sites and investigators is the number of publications released in the specific field. Placing a study in the best recruiting centre, where the appropriate subject population is available, and in a competition-free environment are also critical factors for a successful clinical trial.

The objective of this scenario was to demonstrate how sponsors could identify recruiting sites for participants into clinical trials in the most time- and cost-efficient way. For them, any delay in approval for a successful drug can potentially cost millions in sales, in addition to preventing promising novel therapies from reaching future patients.

2.5.2 Phase IV Post Marketing Surveillance trial

A Phase IV study is a clinical trial, a quasi-experimental study, or an observational study to gather specific information about an approved drug, a biological product, a device, or a procedure. Post-approval research is typically initiated to better understand product use in real-world situations, to obtain evidence for higher reimbursement or submission for expanded labeling, to fulfill a specific requirement of regulatory authorities, or to monitor safety of a drug or device in a larger, non-clinical trial setting.

The need for post-approval surveillance studies stems from inherent limitations in the clinical trial process used for regulatory approval. While studies of safety and efficacy in optimal populations are required to bring a drug or a device to market, it is only through post-approval research and studies of patient outcomes from product use in real-world settings that strong evidence of the effectiveness and safety of a new product emerges. Studies in large populations, with real-world dosing, longer duration of exposure, long-term follow-up, and comparison data based on current physicians' practices, are the most informative approaches to monitoring device and drug safety.

Regulatory initiatives have confirmed the significant value and appeal of information generated from registries that are well-designed and appropriately conducted, analysed, and reported [25]. Observational studies constitute a sub-category of post marketing studies. Observational studies are studies in which the researcher observes and does not alter the participant's experience in any way. In an observational study, treatment decisions, visit schedules, and any tests/measurements are generally left to the discretion of the provider. Observational studies can provide the health community with invaluable data on safety and effectiveness of a product and/or information about the natural history of a disease under standard care practices. These studies provide real-world data that can benefit patients, healthcare providers, pharmaceutical companies, sponsors, and regulatory agencies.

The objective of this scenario was to provide an analytical framework to analyse existing databases' content with information including (but not limited to): demographics, medication used, AEs and medical history. The aim was to detect safety signals as soon as possible, and identify association with factors that may act as causative or predisposing to the AE.

2.5.3 Identification of Relations between Molecular Fragments and Specific Adverse Side Effect Categories (Chemoinformatics)

This scenario focused on the identification of structural features in drugs that may be related to AEs in population sub-groups. The inspiration stems from the application of similar chemical structure-based techniques for the prediction of biological properties including potency, solubility, toxicity, etc. in the drug discovery field.

To succeed in this, we needed detailed information on patients' records including drugs administered and AEs. The availability of this information allowed the application of chemoinformatics methods that could suggest potential chemical structure-adverse effect relationships. A subset of such potential relationships could be subjected to further analysis at a molecular level to elucidate the mechanism of action of the associated molecular sub-structure and the characteristics of the sub-population affected. This required a quantity and quality of information on patients and their characteristics, as well as, on AEs.

The objective of this scenario was to demonstrate the usefulness of Linked2Safety framework to search and identify associations between molecular features (e.g. chemical structure of drugs), and the occurrence of AEs.

2.6 External Evaluations

The second step was to conduct external evaluations. External evaluators were members of the wider scientific community. External evaluations were designed and conducted to eliminate potential bias in the feedback and to gather a wider set of feedback on the utility and functionality of the Linked2Safety platform by people not affiliated with the project. For the external user evaluations, a restricted second version of the integrated Linked2Safety platform was prepared and the installation of all components was done in a controlled environment. For security reasons, the external users did not have access to the full platform and could only analyze synthetic data 1[1] (modeled on real-world data). They were given specific scenarios to evaluate using demonstrations and screenshots, which tested the acceptance of confidence in the Linked2Safety concept and its applicability. The idea behind scenario-based evaluation with these external users is for them to perform a type of task that is typical in their professional work by utilizing the Linked2Safety platform. Despite not having full access to the platform of the integrated real-world data, their insights were expected to have a non-biased view of the Linked2Safety platform, providing reliable and objective results.

2.6.1 Participants

The participants of the external evaluation came from three different groups:

1. medical science analysts,
2. analytic methodology engineers, and
3. data providers,

which coincide with the target groups of users of the system.

Medical science analysts focus their efforts on analyzing data; they rely primarily on using existing statistical or computational methods to test pre-existing hypotheses or to generate new hypotheses depending on the problem on which they are working. They are typically associated with large pharmaceutical industry organizations, academic institutions interested in medical analyses and hospitals and other medical care providers that perform analyses on data as part of their decision support process, prognostics, or other efforts. Medical science analysts routinely seek new sources of data to test their hypotheses with increased statistical power, using standardized analytical tools.

Analytic methodology engineers are focused on developing innovative analytical techniques to perform analyses on data and on evaluating those techniques. They may have a background in statistics, computational intelligence, data mining, software engineering and development, or other fields of study. Their focus is on the development of tools that can either introduce new analytic approaches to solve medical problems through the analyses of medical data or to introduce new versions of analytic methodologies that are expected to have certain advantages over existing ones. Typically, an analytic methodology engineer would utilize the Linked2Safety platform as part of his/her efforts to evaluate newly developed tools and, once the tools are proven to be successful, the platform can also enable quick deployment of his/her work to a large number of medical science analysts for use.

Data providers are institutions that hold medical data; these may be organizations that are responsible for data collected through clinical trials, epidemiological studies, health providers with patients' electronic health records, and others who have the ability to store and use that medical data in some form of research analyses. The primary focus of these users is typically to collect data for scientific research, whilst strictly adhering to legal and ethical limitations.

Table 1 shows the three groups of participants of the external evaluations of Linked2Safety. There were a total of three external evaluation events with a total of 75 participants from the three targeted groups of potential users (35 medical science analysts, 11 methodology engineers and 29 data providers).

Table 1: External Evaluation Participants

Institution	Count	Type
CING	24	Medical science analysts
CING	23	Data providers
CHUV	6	Data providers
CHUV	6	Data providers
UNIMAN	11	Analytic methodology engineers
University of Liverpool	5	Medical science analysts

Each evaluation event typically started with a brief description of the Linked2Safety project aims and scope, which was given as a presentation. The presentation included the results of the internal evaluation (e.g. monetary and time cost of deployment results, findings that replicated scientific knowledge already discovered) this was followed by individual hands-on experience of the Linked2Safety system by participants through three different scenarios (a different scenario for each group) that demonstrated the basic functionality of the system for each group of users' main activities. The workflow instructions were in the form of screenshots on how to use the basic functionality of the Linked2Safety platform. After participants used the system and had hands-on experience with its basic functionalities they completed an evaluation questionnaire (a different questionnaire for each group).

2.6.2 Questionnaire

The Linked2Safety external end-user evaluation questionnaire, which was developed specifically for the purposes of the evaluation of the Linked2Safety system, is structured in five main parts. Part 1 of the evaluation questionnaire refers to users' personal information, such as gender, age, employment and experience. Part 2 allows users to evaluate the following five aspects of the Linked2Safety platform:

1. Analyses space: Questions that fall under the category of analyses space cover issues of subject selection, hypothesis testing, hypothesis generation, data mining, replication testing, time, cost and usability.
2. Linked Data Space
3. Usability
4. Legal and ethical issues
5. Value of the system (for patients, future research)

The third part (Part 3) targets only members of Stakeholder Group 2 (analytic methodology engineers); and the fourth part (Part 4) targets only members of Stakeholder Group 3 (data providers).

Lastly, the fifth part (Part 5) of the external end-user evaluation questionnaire provides users with the ability to express their opinion in a few open-ended questions that focus on ways to improve Linked2Safety.

All questions were in a ‘multiple choice’ format and began with a statement to which the clinical partner was asked to state their agreement by selecting from the options: strongly disagree, disagree, agree and strongly agree. In addition, a ‘not applicable’ option was given should the user feel unable to give an answer. Not all statements were positive (a method commonly used in survey design to ensure that subjects do not attempt to reply completely positively or negatively without closely reading the questions). It is also important to note that in order to avoid bias in participants’ answers (e.g answering with positive replies of either ”agree” or ”strongly agree”), the evaluation questionnaires were completed anonymously.

2.6.3 Questionnaire analysis

Following the data collection of the study, all the data from the different evaluation sessions were input in a statistical package (SPSS) for analysis. For the questionnaire results, descriptive statistics (frequencies) of 75 participants’ answers have been collected to illustrate the users’ perceptions of the Linked2Safety system, broken down into specific categories that can be analyzed in more detail (analyses space, linked data space, usability of the system, legal and ethical issues, value of the system).Associations/correlations between the users’ personal data (e.g. their experience, educational level, age, gender) and their perceptions of the Linked2Safety system (e.g. to what extent they value the system, to what extent they find specific tools user-friendly etc.) were conducted. For these evaluations, a comparison of responses in questions that are common in the three target groups (data providers, medical science analysts and analytic methodology engineers) have been conducted to identify whether there are differences between users' perception of Linked2Safety based on their role.

2.6.4 Participants demographics

As summarized in Table 1, there were three external evaluation events organized as part of the Linked2Safety project that took place in the following partners' premises:

1. CING (with 47 participants),
2. CHUV (12 participants), and
3. UNIMAN (16 participants).

The total number of external evaluators was 75, including 35 medical science analysts, 11 methodology engineers and 29 data providers. 45.3% of participants were male and 54.7% of participants were female. Over 90% of evaluators had either an MSc or PhD. The majority came from medical research institutes (42.7%) or academic institutions (30.7%). With regards to experience, almost half of the evaluators had at least 3 years of experience.

2.7 Results

For all five aspects of the Linked2Safety systems that were examined (analysis space, linked data space, usability, legal and ethical issues and value of the system), the participants' perceptions were overwhelmingly positive. These results are summarized in Table 2, which shows that all three groups had a Mean score between ”agree” and ”strongly agree” (higher than 3.10 out of 4.00 in all cases). Overall, the participants' perceptions of the analysis space were positive (Mean = 3.28, SD(Standard Deviation) = 0.50) as all three groups had a Mean score between ”agree” and ”strongly agree”. The analysis space aspect includes all questions that refer to the overall functionality of the platform, as well as the functionality of the Linked2Safety system in relation to saving time and money when compared with traditional systems.

Table 2: Descriptive Statistics of Participants’ Perceptions in all Five Aspects of Linked2Safety

	N	Mean	Type
Analysis space	75	3.28	0.5
Linked data space	74	3.22	0.49
Usability	75	3.1	0.63
Legal and ethical	72	3.33	0.55
Value of the system	75	3.21	0.43
Valid N (listwise)	71		

Considering the results in more detail, per group of participants, the descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 35) had the lowest scores in the evaluation of the Analysis Space (Mean = 3.22; SD = 0.45). Data Providers (N = 29) had the next highest evaluation score (Mean = 3.31; SD = 0.49) while Methodology Engineers (N = 11) had the highest evaluation score (Mean = 3.49; SD = 0.45). An Analysis of Variance (ANOVA) test showed no significant differences between the three groups (F = 1.423, p = 0.248) and thus no post hoc tests were run. There were also no significant differences among the groups in relation to gender, age, educational and employment.

A one sample t-test was then run to evaluate the inclination of answers (between satisfaction and dissatisfaction), with the value of 2.5 taken as the ‘neutral’ answer to be tested against for each group separately and all the groups together. The aim of this test is to examine whether the participants responses were in general higher than or lower than 2.5 (with 2.5 being the neutral answer between the values of 1 that showed disagreement/dissatisfaction with the particular aspect of the Linked2Safety system that was examined and 4 that showed agreement/satisfaction with the particular aspect of the Linked2Safety system that was examined), in other words it identifies whether the participants' responses were positive or negative at a statistically significant level. The results show a significant positive inclination (p $\leq$ 0.001) in all groups separately as well as together.

At a more detailed level of analysis, we examined the descriptive statistics of individual questions that were part of the analysis space aspect. The vast majority of participants (over 90%) were positive about using the Linked2Safety platform to identify and combine data with other institutions and to locate datasets and subjects to test their hypotheses. Similarly, positive perceptions have been expressed about saving both money and time when deploying a new methodology (Net Per cent Agreement NPA = 90.9%, n = 75) or when locating data and selecting subjects (NPA = 94.3%, n = 75). Seventy per cent (70%) of evaluators agreed that the Linked2Safety platform could increase the statistical power of their experiments. While the participants were generally positive about using data mining to generate further hypotheses (NPA = 91.9%, n = 75), the scenarios they have executed, which used fake data, did not allow them to generate specific results that were worth investigating further, given the dataset used for external evaluation. They were also largely positive about using the Medical Dictionary for Regulatory Activities (MedDRA) for mapping of data (NPA = 91.9%, n = 75).

Furthermore, it was found that for the statement: 'I was able to investigate my hypothesis by testing for associations' only 5.7% disagreed and 94.3% agreed or strongly agreed. For the statement 'The use of Linked2Safety allowed me to successfully test the hypothesis of the study' 13.1% disagreed or strongly disagreed, and 84.7% agreed or strongly agreed.

2.7.1 Linked data space

Overall, the external evaluators' perceptions of the linked data space were positive (Mean = 3.22, SD = 0.49) as all three groups had a Mean score between ”agree” and ”strongly agree”.

Considering the results in more detail, per group of participants, the descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 35) gave a medium evaluation score about Linked Data Space (Mean = 3.23) but with the highest variability in their answers (SD = 0.52). Data Providers (N = 29) gave the lowest evaluation score (Mean = 3.12; SD = 0.47) and Methodology Engineers (N = 10) showed the highest perception score (Mean = 3.33; SD = 0.44), which may be expected given their understanding and utilization of similar services for linking data. All three groups had a Mean score between ”agree” and ”strongly agree”. An ANOVA test performed to investigate differences between the three groups showed no significant differences (F = 0.425, p = 0.656) and thus no post hoc tests were run.

A one sample t-test was run to see the inclination of answers (satisfaction and dissatisfaction), with the value of 2.5 taken as the ‘neutral’ answer: the results show a significant positive inclination ($p \leq 0.001$) in all groups separately as well as together. There were no significant differences among the groups in relation to gender, age, educational and employment.

At a more detailed level of analysis, we analysed the responses to individual questions that were part of the Linked Data Space aspect. The vast majority of answers (over 90%) reflect the perception that the linked data approach developed as part of Linked2Safety could provide a standardized and efficient way to enable merging of data from multiple sources for analysis. In addition, it could provide a meaningful and standardized approach to merging of clinical terminologies across multiple institutions (NPA = 66.7%, n = 75).

2.7.2 Usability

Overall, the external evaluators’ perceptions of the usability of the Linked2Safety system were positive (Mean = 3.10, SD = 0.63) as all three groups had a Mean score between ”agree” and ”strongly agree”. At a more detailed level of analysis, we analysed the responses to individual questions that were part of the system usability aspect. Overall, over 60% of evaluators thought that the platform was easy to use, and that the interface is not complex (NPA = 90.6%, n = 75). Around 80% of participants felt that the analytic space, the mapping tool, and the integration of MedDRA were easy to use, and similarly that the data mining tools were intuitive to use. The majority of evaluators expressed their motivation to use the Linked2Safety platform in the future (NPA = 78.3%, n = 75).

2.7.3 Legal and ethical issues

Overall, the external evaluators' perceptions of legal and ethical issues in relation to the Linked2Safety platform were positive (Mean = 3.32, SD = 0.55) as all three groups had a Mean score between ”agree” and ”strongly agree”. The descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 33) gave the lowest evaluationscore of legal and ethical issues (Mean = 3.24; SD = 0.59). Importantly, Data Providers (N = 29) gave the highest evaluation score (Mean = 3.43; SD = 0.53), while Methodology Engineers (N = 10) had a medium evaluation score (Mean = 3.30; SD = 0.48). Overall, all three groups had a Mean score between ”agree” and ”strongly agree”. An ANOVA test performed to investigate differences between the three groups showed no significant differences (F = 0.914, p = 0.406) and thus no post hoc tests were run.

A one sample t-test was run to see the inclination of answers (satisfaction and dissatisfaction), with the value of 2.5 taken as the ‘neutral’ answer: the results show a significant positive inclination ($p \leq 0.001$) in all groups separately as well as together. There were no significant differences among the groups in relation to gender, age, educational and employment.

At a more detailed level of analysis, we analysed the responses to individual questions that were part of the legal and ethical aspect. Over 90% of evaluators felt that the platform guarantees anonymity of data (through data cubes) (NPA = 93.2%, n = 75), and almost 70% thought that re-identification of individuals was improbable using reasonable financial and technical efforts (NPA = 69%, n = 75).

2.7.4 Value of the system (for patients, future research)

Participants of the external evaluation were aware of the cost deployment results of the internal evaluation of the platform through a presentation of the system that preceded the administration of external evaluation instruments. Overall, the external evaluators’ perceptions of the value of the Linked2Safety platform for patients and future research were positive (Mean = 3.21, SD = 0.43) as all three groups had a Mean score between ”agree” and ”strongly agree”. The descriptive statistics for the responses in the three groups showed that Medical Science Analysts (N = 35) had the lowest evaluation score on the value of the system (Mean = 3.12; SD = 0.43). Encouragingly, Data Providers (N = 29) had the next highest evaluation score on the value of the system (Mean = 3.27; SD = 0.40) and Methodology Engineers (N = 11) had the highest evaluation score on the value of the system (Mean = 3.36; SD = 0.40). Overall, all three groups had a Mean score between ”agree” and ”strongly agree”.

Overall, over 80% of participants felt that the system enabled analyses that maximize the positive effect to the patient (NPA = 80.6%, n = 75). More than 80% of participants valued the version with MedDRA incorporated into the system, while two-thirds thought that the infrastructure would help them examine new analytical techniques, in particular if a large number of organisations are involved. Overall, the vast majority of evaluators (over 85%) thought that the Linked2Safety platform would have an impact on future research, is easy to use (NPA = 75.8%, n = 75) and that they would recommend it to other users and data providers (NPA = 83.8%, n = 75), in particular if it is made publically accessible (NPA = 83.8%, n = 75).

2.7.5 Participants' motivation to become a data provider

Overall, their composite score in these four questions indicates that they had a positive perception towards the idea of becoming a data provider (Mean = 3.38, SD = 0.46).). It is important to note that these results were positive even though participants were aware that there is some cost associated with the decision to become a data provider, which involves the need to employ a data manager, IT manager, legal advisor and project manager for the preparatory activities and a research scientist for the running activities. The group of Data Providers was also asked four specific questions to examine their motivation to become data providers for the Linked2Safety system.

Over 85% of data providers felt that the platform would reduce the time needed for data sharing compared to other approaches and that this should significantly increase the number of samples available (N = 85.2%, n = 75). Around two-thirds of evaluators thought that the process of becoming a data provider is simple (NPA = 72.4%, n = 75) and cost-effective (NPA = 62.9%, n = 75).

2.8 Discussion

From the deployment of the system that included three independent institutions from different European countries we were able to determine the costs for each aspect of the system deployment, as well as evaluate the new scientific potential of the increase in statistical power. It is clear that although only one data provider had whole genome data, while another had candidate gene studies, the overlap in genetic bio-markers as well as phenotypes was significant and allowed for the joint analyses of the datasets with a significant increase in statistical power. This indicated that a wider deployment of Linked2Safety or similar future system will results in significant gains in statistical power, enhancing the discoverability of new knowledge from existing data. Furthermore studies tend to collect a wide array of data beyond their primary and secondary endpoints that may be used either to adjust for known effects, or to study potentially unexpected/unknown effects. By merging data across many studies it's possible to gain sufficient statistical power to discover knowledge related to otherwise rare observations including combinations of biomarkers. Linked2Safety, clearly demonstrates a potential for such a system to enable re-use and a significant increase in data collected as part of isolated studies.A set of external evaluation scenarios were developed for users working in the medical / pharmaceutical industry who are external to the project, in order to reduce possible bias. These were similar to the internal scenarios but, for security reasons, these external users did not have access to the full platform and could only analyze synthetic data. They were therefore given more general, less complicated scenarios to evaluate using demonstrations and screenshots which tested the acceptance of, confidence in and applicability of the Linked2Safety concept.

In all aspects the participants' responses were positive at a statistically significant level, as shown in the one-sample t-tests that were run. The results show a significant positive inclination ($p \leq 0.001$) in all groups separately as well as together, a finding that indicates that the acceptance level is high, their confidence in their ability to deploy the system in the future is high and, lastly their perceptions for the applicability of the concept are positive. In all cases there was no differentiation in participants' responses in relation to the group of users to which they belonged (analysts, methodology engineers or data providers). Further, in all cases, there were no significant differences among the groups in relation to gender, age, educational and employment.

If we look at the results per aspect, with regard to the analysis space, the vast majority of participants (over 90%) were positive about using the Linked2Safety platform to identify and combine data with other institutions and to locate datasets and subjects to test their hypotheses. Similarly, positive perceptions have been expressed about saving both money and time when deploying a new methodology or when locating data and selecting subjects. 70% of evaluators agreed that the Linked2Safety platform could increase the statistical power of their experiments.

With regard to the linked data space, the vast majority of answers (over 90%) reflect the perception that the linked data approach developed as part of Linked2Safety could provide a standardized and efficient way to enable merging of data from multiple sources for analysis. With regard to usability, over 60% of evaluators thought that the platform was easy to use, over 90% of users thought that the interface is not complex and around 80% of participants felt that the analytic space, the mapping tool and the integration of MedDRA were also easy to use. With regard to the legal and ethical issues, over 90% of evaluators felt that the platform guarantees anonymity of data (through data cubes).

Lastly, with regard to the value of the system, over 80% of participants felt that the system enabled analyses that increase the positive effect to the patient; while two-thirds thought that the infrastructure would help them to examine new analytical techniques, in particular if a large number of organizations is involved. The vast majority of evaluators (over 85%) thought that the Linked2Safety platform would impact future research and that they would recommend it to other users and data providers, as the system is currently publically available. It is important to note that participants' perceptions were positive while at the same time they were aware of the cost for deployment results of the platform.

2.9 Conclusion

This chapter presented an innovative and secure semantic interoperability framework that is valuable for pharmaceutical companies, healthcare professionals and patients. Linked2Safety addressed the problem of diversity and complexity of today's legal and ethical regulations imposed at both the national and European legislation level, which make it difficult, risky and expensive to transfer data by sharing EHRs. The proposed solution provided a semantically interconnected approach to sharing aggregate data in the form of data cubes, which eliminated the risks associated with sharing pseudoanonymized (and therefore still personal in some types of data such as genetics) data while enabling the multi-source, multi-type analysis of health data through a single web based secure access platform.

The external evaluation that was conducted put the Linked2Safety theory into practice, allowing both clinical partners and potential external users coming from academia, and the medical and pharmaceutical industry to interact with the system. The research focus of the study was on the documentation of the perceptions of Medical science analysts, Analytic methodology engineers and Data providers on the evaluation of the system with respect to five specific dimensions (analysis space, linked data space, usability of the system, legal and ethical issues, and value of the system).

For all five dimensions of the Linked2Safety system that were examined, the participants' perceptions were overwhelmingly positive, providing evidence of the acceptance of, confidence in and applicability of the Linked2Safety concept. Patient rights are inherently addressed through the design as it minimizes the risk of de-anonymization while still allowing for the mechanisms of the data provider to support removal of subjects from repositories to be propagated to the analyses performed by Linked2Safety. Patients' choices when it comes to personal data protection and supporting scientific advancement frequently involve a delicate balance that is very hard for laymen potential participants to reach a truly informed decision. Through Linked2Safety the risk of privacy breach can be reduced, making it more likely for participants concerned with privacy protection to participate in studies.

Policy makers also have to balance the strictness of legal regulation on the use of personal data so as to both protect the subjects/citizens as well as still enable beneficial research and technological advancement. A wider deployment of such aggregate data based solutions for the joined analyses can enable policy holders to adhere to the strict protection requirements expected by citizens without having a significant impediment to research. However it should also be noted that the proposed approach has focused on specific types of data (categorical, or variables undergoing quantization), to enable wider adoption there is a need to expand future research into these approaches to solve the remaining challenges and enable wider standardization and wider adoption.

Linked2Safety or systems developed in the future based on similar concepts of aggregating data from multiple providers across Europe and beyond in a way that only research focus epidemiological analyses is enabled that adheres to national and international legal and ethical requirements could revolutionize the capacity for knowledge discovery without the need for larger, or significantly costlier studies. The tools exist already to enable standardization of data collected, as well as secure joint analyses; challenges remain however on the political and legal front with ambiguous and clustered legal frameworks. Consent seems to be of vital importance, as there is a lack of standardization on how the use of subject's data is limited to specific application (if at all) making efforts to enable wide ranging aggregate analyses challenging.

Chapter 3

Privacy Preserving Data Publishing of Categorical Data through k-anonymity and Feature Selection [3]

3.1 Introduction

In recent years, with the infiltration of information technology in healthcare, many healthcare related entities such as hospitals and pharmaceutical companies, have vast amounts of patients' data. Although it is clear that sharing such data can increase the likelihood of identifying novel findings or even replicating existing research results, this is not happening due to legal and ethical issues. Attempts have been made by research programs like Linked2Safety [1], for merging data across different healthcare entities and preserve individuals privacy, but one needs to also decide which data to share, so that the likelihood that the shared data will lead to new findings is increased. The problem is more evident in the case of genetic data, since they can reveal the identity of an individual and possible risk factors that s/he may have. Hence they are highly sensitive data and caution is needed even if a part of them is shared.

The aim of Privacy Preserving Data Publishing (PPDP), is to provide the means for publishing data in a way so that the privacy of individuals is preserved with a minimum loss of information [26]. One approach that is employed for data anonymization is k-anonymity [27, 28]. In k-anonymity, a dataset is considered anonymized if the combined values of the *quasi-identifiers* appear at least k times, which means that there is at most $1/k$ probability of identifying an individual using the available data. Features are considered as *quasi-identifiers*, if their combined values can be linked to publicly available information and can lead to the re-identification of individuals [28]. Knowing beforehand all of the features that can be used as *quasi-identifiers* is difficult [28] (e.g. genetic data), hence one could consider all features as such and therefore publish datasets whose records exist at least k times.

The most common methods for achieving the k-anonymity requirement are generalization and suppression [29]. In generalization, the values of features are transformed into more general ones so that details of the individuals are disclosed. For example the address of an individual could be replaced with the zip code, the city or even the country the individual is living, depending on how abstract the information needs to be so that k-anonymity is obtained. This requires the creation of value generalization hierarchies (taxonomy trees) for the *quasi-identifiers*, so that an anonymization algorithm can select the minimum level of generalization needed. Research focused more on the generalization approach, resulting in many such algorithms [30, 31, 32, 33], but their need for taxonomy trees makes it hard for use in high-dimensional data. Algorithms that can cope with the lack of user defined taxonomy trees such as Top-Down Refinement (TDR) [34] exist, but in general, high-dimensionality can severely affect the information contained in the anonymized datasets [35].

In suppression, part of the dataset is removed so that the k-anonymity requirement is not violated. This can be achieved by various suppression methods, like record suppression and attribute suppression. For example in record suppression, all records (instances) that appear less than k times in the dataset are removed, whereas in attribute suppression, features that have values that do not conform to k-anonymity are removed. This approach can lead to substantial information loss since data are removed instead of being replaced by generalized values, hence caution is needed when used. Methods that employ suppression for obtaining k-anonymity before data publishing have been proposed [28, 36, 37] and there has been some focus on suppression methods that also combine feature selection [38, 39]. With the use of feature selection, the dimensionality of the data is reduced and features that are not related to the problem of interest can be removed. With fewer features, the probability of having records that are unique decreases, thus less suppression is required for obtaining k-anonymity.

In this chapter, k-anonymity through Pattern Based Multidimensional Suppression (kPB-MS) is proposed for PPDP. The algorithm uses feature selection for reducing the data dimensionality and then combines attribute and record suppression for obtaining k-anonymity. The proposed algorithm can be used on categorical data for classification tasks. The structure of the rest of the chapter is as follows: Section3.2 introduces a new measure which is used in the proposed feature selection algorithm presented in Section 3.3. In Section 3.4 the anonymization methodology is shown, whereas Section 3.5 presents the datasets and the evaluation methodology used. The last three sections contain the results of the algorithm along with a discussion and the concluding remarks.

3.2 Pattern Based Classification Accuracy (PBCA)

In this section a new measure is proposed, which will be used as the performance metric of the feature selection step of the anonymization methodology. PBCA uses the discrimination power of the patterns in a dataset to calculate the accuracy that can be obtained by classifiers. It is applicable on categorical data, and considers each input instance as a pattern. As will be shown it is the upper limit of the classification accuracy that can be obtained by classifiers when the entire dataset is used both for training and testing. In the following table, the mathematical notations that will be used in the rest of the chapter are given.

Table 3: Mathematical notations

Notation	Meaning
x	Single value variable
$\mathbf{x}$	Vector
$\vert\mathbf{x}\vert$	Number of elements in vector **x**
X	nxd Matrix
$X_{i,j}$	The value of the *j*-th feature of the *i*-th instance

The PBCA of the features in X and the response variable **y**, can be calculated using the following equation

$$PBCA(X, \mathbf{y}) = \frac{\sum_{i=1}^{p} max(T_{i,c} : c \in [1, |\mathbf{c}|])}{n} \quad (1)$$

where p is the number of rows of the contingency table (number of unique patterns), T is the contingency table, $T_{i,c}$ is the number of instances of row/pattern i in the contingency table T that belong to class c, $|\mathbf{c}|$ is the number of classes of the response variable and n the number of instances in the dataset.

PBCA is model free and can capture both linear and non-linear dependencies among the features and the response variable. The complexity of the approach for calculating the PBCA is $O(n)$, since it requires one pass from the data to create the contingency table and a single pass from the contingency table to calculate the PBCA.

A disadvantage of the measure is that it is biased towards variables with many categorical values. For example, if the unique ID of each instance is included in the dataset, then the PBCA would be 100% regardless of the values of the rest of the variables.

Table 4 shows a contingency table of a hypothetical dataset. The dataset has three features (f_1, f_2, f_3) with three categorical values each, whereas the response variable (Class) is binary. The first row indicates that all 10 instances with the values (pattern) $\{1,2,1\}$ for features f_1, f_2 and f_3 respectively belong to class ”0”. Similarly the rest of the rows indicate how many instances for each pattern belong to each class in the dataset. Such a contingency table can be created with a single pass from the dataset using a hash table, in which the key used is the pattern of the instances.

Table 4: Example of a contingency table

f_1	f_2	f_3	Class=0	Class=1
1	2	1	*10*	*0*
1	2	2	*0*	*10*
2	3	1	*10*	*20*
2	3	3	*10*	*10*
3	1	1	*20*	*0*
3	1	3	*0*	*10*

When applying (1) on the dataset in Table 4, a PBCA of 80% is obtained, which means that $(10 + 10 + 20 + 10 + 20 + 10) = 80$ instances out of the total 100 would be correctly classified.

An ideal classifier would classify an instance to class ”0” if most of the instances of a pattern belonged to ”0”, and ”1” otherwise. Using the dataset shown in Table 4, it would classify instances that match the pattern of the first row as class ”0”, instances that match the pattern of the second row as class ”1” and so on. In the case of a tie among the number of instances that belong to each class of a pattern, any of the classes can be selected. Hence by selecting the most probable class for each pattern, as observed in the dataset, the PBCA for the dataset can be calculated. As can be seen, since the most probable class for each pattern is selected, the PBCA indicates the upper limit of the classification accuracy that can be obtained when the entire dataset is used for both training and testing by a classifier.

3.3 Pattern Based Feature Subset Selection (PB-FSS)

As mentioned, with feature selection one can reduce the dimensionality of a dataset and remove any redundant and non-informative features. For the needs of the proposed anonymization process, a new feature selection algorithm is presented and shown in Algorithm 1. PB-FSS has three main steps. The first step is to perform Forward Sequential Feature Selection (FSFS) using PBCA as its performance metric. The second step is to remove any features that are redundant in the features subset and the final step is to order the selected features based on their importance.

Algorithm 1 PB-FSS

```
Require: $X, \mathbf{y}$
$\mathbf{s} \leftarrow FSFS(X, \mathbf{y})$
$\mathbf{s} \leftarrow removeRedundantFeatures(\mathbf{s}, X, \mathbf{y})$
$\mathbf{s} \leftarrow sortFeatures(\mathbf{s}, X, \mathbf{y})$
return s
```

FSFS begins with an empty features subsets and sequentially adds the features that increase the performance metric. At each iteration of the FSFS, the feature that maximizes PBCA when added to the selected features subset is found. If the selected feature increases the PBCA when included in the features subset **s**, the process is repeated. If the PBCA is not increased then the process finishes and the feature subset without this feature is returned. Since the entire process depends on the PBCA, to overcome its bias towards features with many categorical values, cross validation is used in the PBCA calculation.

When adding features in the subset using FSFS, it is possible that some of the already selected ones become redundant. For example a new feature that is selected, interacts with one or more features in a way that the effect of a previously selected feature is masked and hence is no longer needed. In such cases, features that become redundant need to be removed. The second step of PB-FSS is responsible for removing such features and is performed in the function *removeRedundantFeatures*. At each iteration of the function, a feature is removed from the subset and the PBCA is recalculated. If the PBCA remains the same after the removal of a feature, that feature is considered as redundant and it is removed from the selected features subset, otherwise it is kept in the final subset.

The third step of the PB-FSS, sorts the selected features in descending order based on their contribution to the PBCA. To calculate the PBCA contribution of a feature the following procedure is followed. Each feature is removed from the subset to calculate the PBCA that can be obtained without it and then it is added back to the subset. The feature that produced the smallest PBCA difference when removed, is considered as the one with the smallest contribution and is added to the beginning of a new subset list. This is repeated until no features remain in the initial subset list. The reason the sorting is performed in a backward feature selection manner, is that this allows to take into consideration feature interactions and produce a better ranking.

Since PB-FSS is based on the PBCA, which is model free, it can capture any type of non-linear interactions among the features. Additionally, the algorithm is able to remove features whose effect can be expressed by the interactions of the other selected features, which can further decrease the size of the returned subset. Finally, it provides a ranking of the features that accounts any feature interactions that exist. A disadvantage of the method, due to its forward selection strategy, is that it could end up in not selecting interacting features with low main effects, such as in the XOR problem.

3.4 k-anonymity through Pattern Based Multidimensional Suppression (kPB-MS

Data anonymization using the suppression method of k-anonymity, can result in the removal of a substantial amount of the initial instances, especially when the dataset is high dimensional. Additionally, by removing instances from the dataset there is the risk of removing important information, which can lead to poorer classification results or even to wrong models. To address these issues, a new anomyzation process called kPB-MS is proposed.

Algorithm 2 kPB-MS

```
Require: $X, \mathbf{y}, k, t$
$\mathbf{s} \leftarrow PB\text{-}FSS(X, \mathbf{y})$
while $1$ do
    $Z \leftarrow anonymization(X, \mathbf{y}, \mathbf{s}, k)$
    $loss \leftarrow instanceLoss(X, Z)$
    if loss $>$ t then
        $\mathbf{s} \leftarrow removeLastFeature(\mathbf{s})$
    else
        $break$
    end if
end while
return $Z$
```

The steps of kPB-MS are shown in Algorithm 2. Initially PB-FSS is performed for reducing the dimensionality of the dataset and for ranking the selected features based on their importance. In the anonymization step, a contingency table with each pattern in the reduced dataset is created like in Table 4. For each pattern, a new 2x|**c**| contingency table is created, where |**c**| is the number of classes. An example of such a contingency table is shown in Table 5.

Table 5: Contingency table for testing the effect of k-anonymity

	Class=0	Class=1
initial	*10*	*20*
k-anonymized ($k = 15$)	*0*	*20*

The first row has the initial number of instances of each class for that pattern. The second row contains the remaining number of instances once k-anonymity is performed. This means that in the second row, the number of instances of the classes that violate k-anonymity are replaced with the value of zero. Fisher's exact test is then performed on this contingency table, for testing the significance of the change in the number of instances in each class. In case the change is statistically significant, all of the instances of that pattern are removed, instead of only removing the ones that do not conform to k-anonymity.

Finally, the percentage of the lost instances ($loss$) from the anonymization process is calculated. In case the $loss$ is above the user's defined threshold (t), the least important feature (in this case the last feature in the subset) is removed and the process is repeated, otherwise the anonymized dataset Z is returned.

3.5 Evaluation Methodology

In this section a description of the datasets used is given along with the preprocessing steps applied on them. Then the process followed for the analysis of the datasets is described.

3.5.1 Datasets

A summary of the datasets used in the evaluation is given in Table 6. All of the datasets are from the UCI Machine Learning Repository [40], except HTS which was created from PubChem[1] . Since the proposed algorithms require categorical data, any continuous features in the datasets were discretized using the Minimum Description Length Principle (MDLP) [41] algorithm.

Table 6: Summary of the datasets

Dataset	#Features	#Instances	#Classes	Class Instances
RETINOPATHY	19	1,151	2	540 / 611
SPECT	22	267	2	55 / 212
SPLICE	60	3,190	3	767 / 768 / 1,655
HTS	1,024	3,115	2	2,000 / 1,115
DOROTHEA	100,000	1,150	2	112 / 1,038

[1] http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=633

The datasets are from different areas of life sciences. RETINOPATHY has features extracted from the Messidor image set [42] to predict whether an image contains signs of diabetic retinopathy or not, whereas SPECT contains data on cardiac Single Proton Emission Computed Tomography (SPECT) images and is on the classification of normal and abnormal patients. SPLICE contains gene sequences and is on the identification of splice junctions. SPLICE is the only dataset with three classes. HTS is on drug discovery and was created using the compounds and outcomes from PubChem. The fingerprints were calculated using the LiSIs platform [43] using a radius equal to two and a fingerprint size of 1024. The initial distribution of the classes for this dataset was highly imbalanced (99% / 1%), thus resampling was used for selecting a subset of the instances of the frequent class (2,000 instances). DOROTHEA is also on drug discovery and it is the largest dataset used in the experiments. This high-dimensional dataset is imbalanced, with its less frequent class representing 11% of the instances.

3.5.2 Analytical Methodology

For the evaluation of kPB-MS, the following methodology was followed. Initially each dataset was anonymized using kPB-MS. Different k values for k-anonymity were used and the accepted loss of instances (t) was set to 10%. Specifically k was set to the values 1, 3, 5, 10 and 20. For the case that k was set to 1, the non-anonymized dataset that results from PB-FSS, is used, so that the effects of k can also be observed. All of the features in the datasets were considered as *quasi-identifiers*.

Once the anonymized datasets were produced, classification with a 10-fold stratified cross validation was performed on the full and the anonymized datasets. The same folds were used in the training and testing of both the full and the anonymized datasets. Specifically, the test folds were always the same among the full and the anonymised datasets, but the training folds could either be the same or use a subset of the instances in the anonymized datasets. This is due to the fact that instances could be removed during the anonymization process. The statistical significance among the results was calculated with a paired t-test and it compared the accuracies on the 10 test folds of the anonymized and the full datasets. The difference was considered as statistically significant, when the obtained p-value was less than 0.05.

For the classification task, four non-linear classifiers were selected. The Support Vector Machine (SVM) [44] with its default parameters, the 1-Nearest Neighbour (1-NN) algorithm [45], the C4.5 decision tree [24] with its default parameters, and the Random Forest (RF) algorithm [23] using 50 trees. The RF and SVM were selected because they were found having the best performance in an extensive test of multiple algorithms for classification [46] and are also widely used in the medical domain. Additionally they represent different learning families, the ensemble and statistical learning family respectively. To see how kPB-MS performs in a non-ensemble method, the C4.5 algorithm was chosen since it can also be directly compared with the RF that uses multiple decision trees. Finally the 1-NN algorithm was selected because the PBCA measure has similarities to the algorithm and it would be of interest to see if 1-NN performs well with kPB-MS. Their accuracy were calculated in R using the RWeka package [47] for k-NN and C4.5 and the e1071 package for the SVM and the randomForest package [48] for RF.

3.6 Results

For each classifier, its results on the full and anonymized datasets are illustrated in Figure 4. The x axis of each plot represents the value used in k-anonymity, whereas the y axis represents the classification accuracy obtained. The colors of the boxplots indicate if the classification accuracy using the anonymized dataset had a statistically significant change compared to the one obtained when the full dataset is used. White indicates no significant accuracy differences, green indicates that better accuracies were obtained with the anonymized dataset and red indicates that better accuracies were obtained with the full dataset.

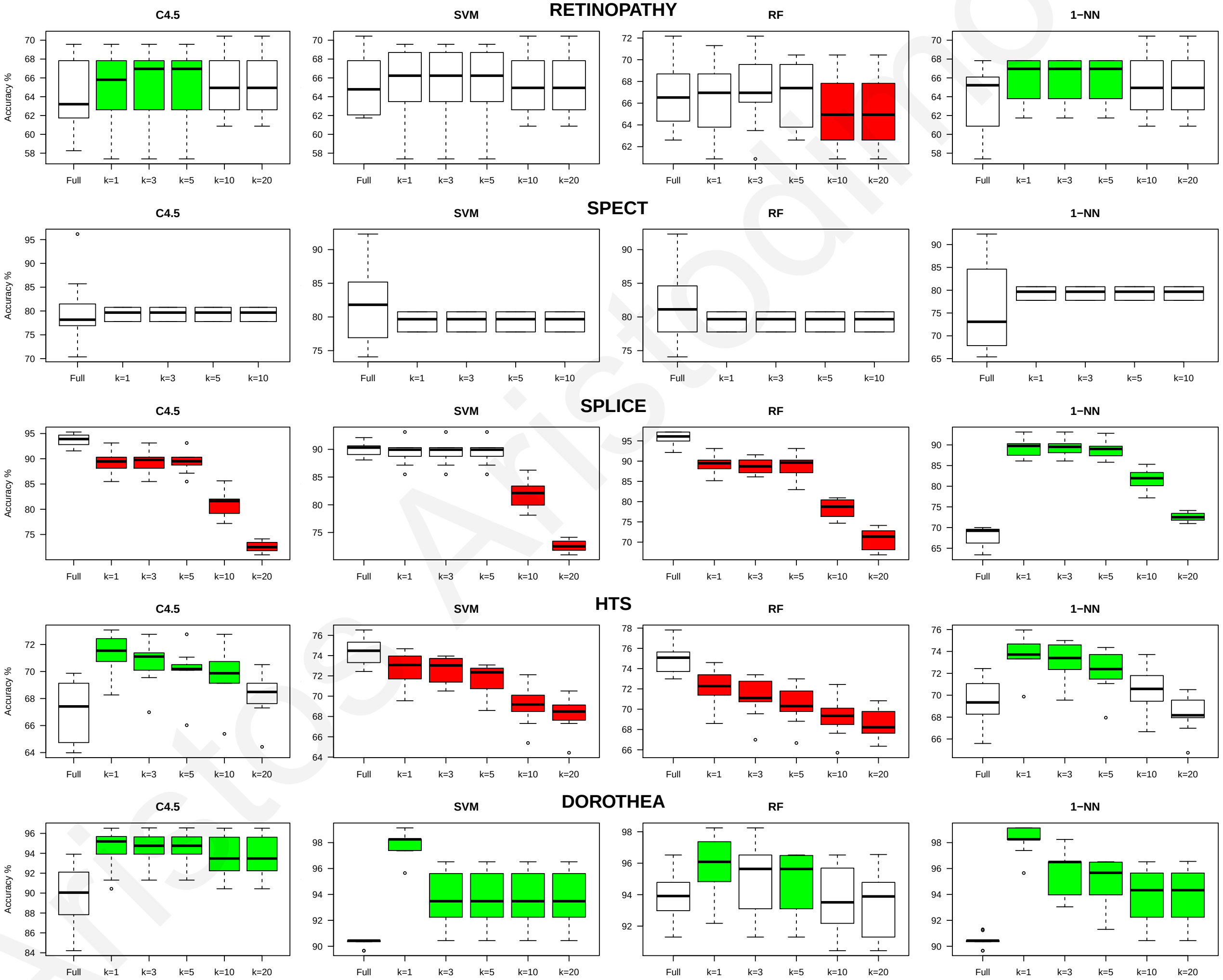


Figure 4: Results on the full datasets and the kPB-MS anonymized datasets with different values for *k*

Table 7: The effect of k on the dimensionality and the instance loss

	Features used (Instance loss)				
Dataset	**k=1**	**k=3**	**k=5**	**k=10**	**k=20**
RETINOPATHY	4 (0%)	4 (3%)	4 (5%)	2 (2%)	2 (5%)
SPECT	1 (0%)	1 (0%)	1 (0%)	1 (0%)	0 (100%)
SPLICE	4 (0%)	4 (2%)	4 (8%)	3 (1%)	2 (6%)
HTS	25 (0%)	25 (7%)	21 (10%)	13 (9%)	8 (9%)
DOROTHEA	19 (0%)	19 (4%)	19 (4%)	19 (8%)	14 (10%)

In Table 7, the features retained in the anonymized dataset and the instance loss for different values of k is shown. For the SPECT dataset, with $k = 20$, the removed instances were more than the accepted threshold so no data could be published. For the rest of the datasets, anonymized data could be published for all of the tested values of k.

As can be seen in Figure 4, the PB-FSS datasets with $k = 1$ (in this case we only notice the feature selection effect of PB-FSS) did not affect negatively the classifiers accuracies in 80% of the cases. Hence the proposed feature selection algorithm is capable of selecting informative features. In the cases that it did affect negatively the accuracies, it is probably caused due to the removal of features by PB-FSS that interacted with others and had low main effects. This is due to the forward selection strategy the algorithm uses.

With the increase of the value of k, the obtained accuracies are affected as expected [38, 49]. This is due to the fact that more instances are being suppressed, which can cause the use of fewer features when the instances loss is above the pre-specified threshold. The effect is more visible for values larger than 5 in the datasets tested (see Table 7). This also shows that before publishing data an analysis on the effect of different values of k should be performed before selecting the most appropriate value, so that a balance between privacy and information loss is acquired.

In general kPB-MS with a $k > 1$ did not have a statistically significant negative effect on the obtained accuracies in 74% of the test cases. The best results were obtained in SPECT and

DOROTHEA. In SPECT only one feature was selected and it had enough information to produce similar accuracies with the full dataset. In DOROTHEA, the anonymized datasets had better or similar accuracies with the full dataset since the high dimensionality was affecting the classifiers ability to model the data.

3.7 Discussion

The proposed kPB-MS, combines feature selection with both attribute and record suppression and additionally provides a method for selecting the accepted percentage of records to be lost. The algorithm attempts to reduce the information lost and as seen from the results it accomplishes this in most of the tested datasets.

The first step of kPB-MS uses the PB-FSS, hence informative features are selected and ranked according to their importance. This can help researchers focus on these features and help them interpret their results. Additionally by sharing the anonymized datasets produced by kPB-MS they increase the probability of replicating the results obtained since they are focusing on informative features.

Since it is using suppression, no user defined taxonomy trees need to be defined as is the case with generalization techniques [30, 31, 32], which makes the algorithm easier to use by non-domain experts. Additionally, one does not need to know beforehand, which features are the possible *quasi-identifiers*, since all of the features can be treated as such. This can further reduce the risk of publishing data that might accidentally contain *quasi-identifiers* that do not adhere to the k-anonymity requirement.

kPB-MS uses multidimensional suppression, which means that it is taking into consideration all of the value combinations of the features for preserving k-anonymity. A similar approach is also followed by kACTUS [38]. kACTUS uses C4.5 for building a classification tree and uses the selected features of the tree to rank the features and decide which to suppress when they do not comply to k-anonymity. Due to the use of C4.5, kACTUS is bound to the performance of the classifier and in high dimensional data like DOROTHEA, kPB-MS is expected to perform better.

In [39] DMPD is proposed, which uses a genetic algorithm for searching for optimal feature set partitioning. One difference with kPB-MS is that it produces multiple feature subsets, which can all be used by classifiers and then combine their classification predictions. This approach is expected to produce less suppressions than kPB-MS, since it can create many subsets with a small number of features and hence have a smaller probability to not adhere to k-anonymity. On the other hand it might be computationally demanding when used on data which are high dimensional like genetic data, due to the increased search space and the need for parameter optimization. kPB-MS would require running PB-FSS on the dataset once to get a ranked list of informative features and then the optimization of k and the instance loss threshold t can be performed on the reduced dataset. Hence it is less computationally demanding since the optimization of the parameters does not require rerunning the feature selection process on the entire dataset each time. Additionally PB-FSS is also computationally efficient, since it is based on forward sequential selection and its performance metric (PBCA) can be calculated in O(n).

To the authors knowledge, there are no similar studies on the datasets used in this chapter. Both DMPD [39] and kACTUS [38] were shown to have significantly better results than traditional methods, but were not compared with each other. Since the performance of kACTUS is bound to C4.5, it is expected that if evaluated on the five databases investigated in this study, its performance would be comparable to the C4.5 results documented in Figure 4 (for the full datasets). Based on this, it can be inferred that kPB-MS is expected to give similar or better results than kACTUS for 4 out of the 5 databases investigated (as documented in Figure 4 for SPECT, RETINOPATHY, HTS and DOROTHEA).

The selection of the thresholds of k and the instance loss threshold t in kPB-MS affect the remaining number of features and instances of the dataset after the anonymization. The biggest effect is from the value of k, since the larger the value the more features get suppressed to have at least k instances per pattern that exist in our dataset. But in the case that the dataset is large and one does not mind losing a big percentage of the instances (e.g. if there were 100 thousand instances, then by setting t to 50% would still leave us in the worst case with a good number of instances for most problems). By setting t to a lower value it is possible to preserve more features. This would be of interest in case we expect that the problem that we want to solve with the specific data is dependent on a large number of features and thus by removing many features we would negatively affect the accuracy of any trained model. Thus, k needs to be set based on the restrictions one has regarding the privacy level that has to be achieved, whereas t can be selected based on the number of instances available and whether one is interested in preserving a larger number of features or not.

3.8 Conclusions

A new method for k-anonymizing datasets has been proposed. This includes the proposal of a new measure and a new feature selection algorithm along with multidimensional suppression. The proposed measure (PBCA), is model free and has a complexity of O(n), which reduces the computational demands of the feature selection algorithm and makes the whole process applicable for high-dimensional data. With PB-FSS, informative features are selected and ranked so that only such features are shared. This can further increase the probability of sharing data that can lead to replicating results.

The multidimensional suppression procedure of kPB-MS takes into consideration the instances removed and the effect of record suppression on classifiers. The first is obtained by allowing the user to define the accepted loss of instances, whereas the second is obtained by testing the significance of the suppression using Fisher's exact test on each pattern. As shown in the results, the algorithm did not negatively affect the classifiers in 80% of the test cases, indicating that kPB-MS can be used in privacy preserving data publishing.

As seen, the value of k can affect the overall results, hence before publishing the data, a proper value for k should be selected. The selection should be made in a way that it provides a balance between the obtained privacy and the information lost. For this task, the evaluation methodology used here can be used on different values of k for guiding such a decision.

Since PB-FSS is using a forward selection method, it can miss interacting features with low main effects such as in the XOR problem. Thus, as part of future research, different search strategies will be tested for overcoming this disadvantage. One such approach could be the use of nature inspired search strategies. Furthermore, work will be done in the expansion of the algorithm for being directly used on continuous features.

Chapter 4

A Supervised Density Based Discretization Algorithm for Big Data Classification Tasks in the Medical Domain

4.1 Introduction

Discretization is a preprocessing step in which continuous values of features are transformed into categorical. Specifically, in discretization, a finite number of intervals is provided and values that lie between two consecutive intervals are replaced by a value which characterizes that range. This process can also be considered as a dimensionality reduction methodology, since the initial spectrum of a feature's values is reduced to a smaller finite set of values [50, 51]. This chapter focuses on supervised discretization, and specifically on its application as a preprocessing step for classification tasks.

The necessity of discretization is due to the fact that a number of existing classifiers and statistical tests, rely on having only discrete data as input. For example C4.5 [24] has an embedded method for discretizing data, whereas if someone wanted to test for association between weight and diabetes using Pearson's Chi Square Test, then the weight would have to be converted into discrete values so as to be able to create the contingency table for the test. Additionally, with discrete data, the processing time of a classifier can be reduced [52] and the results can be easier to interpret, whereas the use of a discretized version of the data can also improve the obtained accuracy by a classifier [50, 53, 54]. With the big data era, discretization algorithms need to be able to handle such data efficiently without compromising their performance. Hence it is important to have discretization algorithms that are computationally efficient and can harness the power of high performance computing, while minimizing the information lost.

4.1.1 Related Work

Discretization algorithms can be categorized based on some of their main characteristics [50]. A main characteristic is if it is using the class label (*supervised*) or not (*unsupervised*), when deciding for the intervals. If a discretizer is independent of a learner then it is *static*, otherwise it is *dynamic*. Algorithms that discretize each feature separately are called *univariate* and if they consider all attributes *multivariate*. In case only a part of the information of a feature is used when deciding for an interval, the algorithm is considered as *local*, otherwise as *global*. Another characteristic is based on whether the intervals are selected simultaneously (*direct*) or one at a time (*incremental*). The last two main characteristics of discretizers are the evaluation measure used to select the most appropriate interval and the procedure used to create the intervals (*splitting* or *merging*). In the splitting approach, a discretizer starts with an empty set of intervals and each added interval divides the domain, whereas in the merging approach it begins with a set of intervals which are being merged based on a criterion. In case the algorithm is incremental, and uses the splitting approach then it is called Top-Down, whereas if it is using merging, then it is called a Bottom-Up approach.

4.1.1.1 Supervised Discretization

In supervised discretization, an algorithm tries to transform each feature from continuous to categorical, by using the class label of each instance. By transforming continuous data to discrete, there is the risk of information loss, especially if a very small number of intervals is created. Hence, a discretization algorithm needs to find a balance between the number of intervals created and the information lost.

In [41], Fayyad and Irani presented an information theory based discretizer, known as Minimum Description Length Principle (MDLP). Its name is from the criterion used for deciding when to stop creating more intervals and uses a Top-Down approach. The values of a feature are initially sorted, and each midpoint between two consecutive values is considered as a new interval. Then it recursively selects the interval that produces the minimum class information entropy, until the stopping criterion is satisfied. A distributed version of MDLP has been proposed [55], allowing it to be used on big data. Numerous information theory based algorithms have been proposed in the literature [56, 57, 58] but MDLP is one of the most widely used discretizers.

ChiMerge [59], is a Bottom-Up discretizer that uses Pearson's Chi Square test (χ^2) to decide if two intervals need to be merged or not. It initially sorts the values of a variable and adds each value in a separate interval. Then it selects the two adjacent intervals with the lowest χ^2 value and merges them. This is repeated until no adjacent intervals have a χ^2 value below the predefined threshold. Again, many algorithms were proposed based on χ^2 [60, 61, 62] and were found to have good results in [50].

Another Top-Down method is Class-Attribute Interdependence Maximization (CAIM) [63]. This algorithm also sorts the values of a feature and then considers the midpoints between the values as possible intervals. It then repeatedly selects the midpoint that maximizes the class-attribute interdependence as a splitpoint. If the current best value is smaller than the global maximum, the procedure stops. CAIM produces at least *C* intervals, where *C* is the number of classes. A version of CAIM that can better handle imbalanced datasets was also proposed [64], whereas a version for handling multi-label data is also available [65].

4.1.1.2 Density Based Discretization Algorithms

The density of a feature can be estimated using parametric or nonparametric methods. Parametric methods are more appropriate when the distribution of the data of a feature follows a known distribution. For example if the normal distribution is assumed, then a parametric method can be used to calculate the mean and variance of the distribution through maximum likelihood estimates [66].

Nonparametric methods are used when the distribution of a feature is unknown, since they can fit different density forms. One such method is kernel density estimation, in which a kernel is used with a smoothing parameter for estimating the density of a feature. A simpler method for visualizing such data, is the use of histograms, in which the values of a feature are represented by bins and the density of the feature in each bin is shown by its height.

Two well known discretization algorithms that use binning, are the equal-width and equal-frequency discretizers. These are unsupervised methods, since they do not take into consideration the class label of the data. Equal-width takes as input the wanted number of bins (*k*), and creates *k* equally sized bins. In equal-frequency, *k* bins with approximately the same number of values are created. These two algorithms did not have a good performance in general in the tests performed in [50]. A more sophisticated unsupervised density based discretization algorithm is TUBE [67], which uses the log-likelihood and cross-validation to select the intervals and decide when to stop splitting the data. It was shown that it can better estimate the density of the data compared to equal-width and equal-frequency in the majority of the tests, where variables had less than 20 unique values, but it is not as fast as equal-width and equal-frequency.

4.1.2 Contribution

The main contribution of this work, is that it turns the fast unsupervised density based discretization into supervised, while having a performance similar to the state of the art supervised discretization algorithms. It initially uses the simple binning approach of the equal-width discretization, along with a statistical test to merge any consecutive bins that have a similar density on each class. Then the resulting intervals of each class are merged to produce a supervised discretization that is computationally efficient and hence suitable for big data.

4.2 Methods

4.2.1 Supervised Density Based Discretization

The proposed Density BAsed Discretization (DBAD) algorithm is based on the hypothesis that intervals with different densities can be used to characterize a class, and essentially be used to help in separating instances of different classes. For example, Class 1 might have a higher density in the range [0,1], whereas Class 2 in the range of [0.5,2]. Based on this, the following intervals of interest can be created {[0, 0.5], (0.5, 1], (1, 2]} to discretize the data. It is also possible that low density intervals of a class can help separate it with another class that has no instances in those ranges, hence one should not only focus on identifying the high density intervals of each class, but instead should identify the intervals in which densities change.

DBAD is a heuristic algorithm that tries to emulate the way a human would identify similar density intervals with the help of a histogram. Figure 5 illustrates the process followed by DBAD for a class on a feature. It begins with equal-width binning (histogram (a)) for estimating the

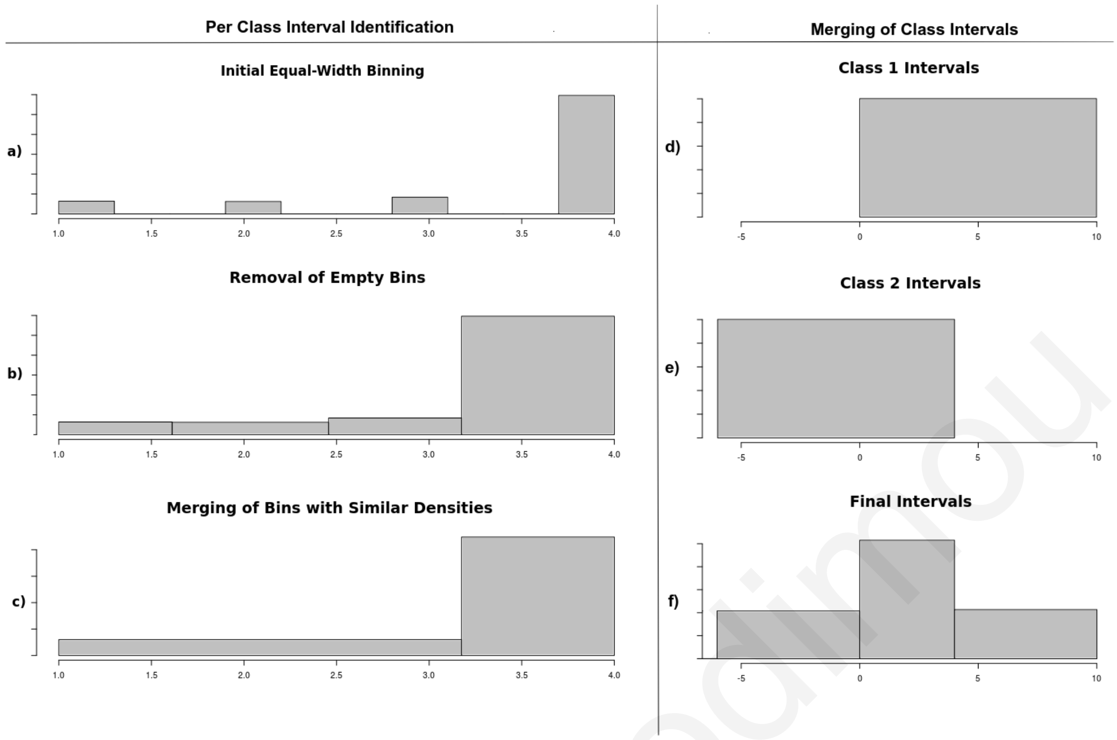


Figure 5: Interval identification for a single class of a feature using DBAD (left) and merging of the intervals of each class for identifying the final intervals of a feature (right). The top left histogram (a) illustrates the initial equal width binning. The histogram in middle left (b) illustrates the effect of the removal of the empty bins. The bottom left histogram (c) shows the final bins after merging any consecutive bins with similar densities. The top right (d) histogram shows the intervals selected for Class 1 whereas underneath it (e) the intervals of Class 2 and the bottom right histogram (f) the final intervals after the ones of the two classes got merged.

density of a feature in different intervals (bins). Each of the created bins contains a number of instances, which defines the density of the feature in those intervals. The more instances in a bin, the higher its density. Once the bins are created, DBAD removes any empty bins by allocating their range to their neighboring non-empty bins (histogram (b)). Then it identifies any consecutive bins with similar densities, using a statistical test, and merges them (histogram (c)). This way, DBAD ends up with bins that have different densities.

This procedure is performed on the instances of each class for a feature, so as to have a supervised approach, which will further help with classification tasks. Once the final intervals for each class are identified, they get merged to get the final discretization intervals. An example of this final step of DBAD can be seen in histograms on the right in Figure 5. As illustrated, Class 1 intervals are in the range [0,10] (histogram (d)), whereas for Class 2 in the range [-6,4] (histogram (e)), so the final intervals for this feature are {[-6, 0], (0, 4], (4,10]} (histogram (f)).

Algorithm 3 DBAD

```
Require: feature values $\mathbf{x}$
Require: unique class values $\mathbf{y}$
 1: featureBins ← $\emptyset$
 2: for each $y \in \mathbf{y}$ do
 3:     $k \leftarrow \lceil \log |\mathbf{x}^y| \rceil + 1$
 4:     bins ← $createEWBins(\mathbf{x}^y, k)$
 5:     bins ← $removeEmptyBins$(bins)
 6:     bins ← $mergeSimilarDensityBins$(bins)
 7:     featureBins ← featureBins $\bigcup$ bins
 8: end for
 9: featureBins ← $removeEmptyBins$(featureBins)
10: return featureBins
```

The steps followed by DBAD on a feature, are shown in detail in Algorithm 3. DBAD gets as input two vectors, which contain the values of the feature that will be discretized (**x**) and the unique class values of the dataset (**y**). To better capture the intervals in which densities change, a proper number of bins (k) needs to be selected. DBAD uses Sturge's rule [68] (step 3 in Algorithm 3) for this and then creates k equal-width bins with the function *createEWBins*.

With equal-width binning, it is possible to end-up with bins that have no instances in them. To eliminate such bins, *removeEmptyBins* is used (see Algorithm 4). Initially, it merges consecutive empty bins, by replacing the right interval of the left empty bin, with the right interval of the right empty bin (step 7), and then removing the latter from the bins (step 8). Once the consecutive empty bins are merged, the new empty bin needs to be removed as well. To perform this removal, it is needed to assign its intervals to its surrounding bins. Initially, a new split point (w) is calculated (step 11) using (2)

$$w \leftarrow \boldsymbol{bins}_z^{left} + \frac{\boldsymbol{bins}_{z-1}^n}{\boldsymbol{bins}_{z-1}^n + \boldsymbol{bins}_{z+1}^n} * (\boldsymbol{bins}_z^{right} - \boldsymbol{bins}_z^{left}) \tag{2}$$

Algorithm 4 removeEmptyBins

```
Require: bins
 1: $i = 1$ #indexing starts at 0
 2: while $i < |bins|$ do
 3:     if $bins_i^n == 0$ then
 4:         $z = i$
 5:         $i = i + 1$
 6:         while $bins_i^n = 0$ do
 7:             $bins_z^{right} = bins_i^{right}$
 8:             $bins = bins - bins_i$
 9:             $i = i + 1$
10:         end while
11:         calculate $w$ using (2)
12:         $bins_{z-1}^{right} = w$
13:         $bins_{z+1}^{left} = w$
14:         $bins = bins - bins_z$
15:     else
16:         $i = i + 1$
17:     end if
18: end while
19: return bins
```

where $\boldsymbol{bins}_{z-1}^n$ and $\boldsymbol{bins}_{z+1}^n$ represent the number of instances of the bins on the left and right of the empty bin, whereas $\boldsymbol{bins}_z^{left}$ and $\boldsymbol{bins}_z^{right}$ are the left and right intervals of the empty bin that will be removed. Then the right interval of the bin on the left, and the left interval of the bin on the right are replaced by w (steps 12-13) and the empty bin is removed (step 14). This approach, ensures that the remaining bins cover the range of the removed bin based on their density, thus bins with more instances (more dense) get a larger range of values. This effect can be seen in Figure 5 and the plots (a) and (b). For example the first two non empty bins, which have a similar density, shared the empty bin's space between them, whereas the last bin got the largest share of the empty bin that was on its left, because it was much more dense than the third non-empty bin.

The next step of DBAD is to merge any consecutive bins with similar densities (similar number of instances) using *mergeSimilarDensityBins*. To decide if two consecutive bins have a similar density, a metric is needed which will be less strict with bins representing a small number of

Algorithm 5 mergeSimilarDensityBins

```
Require: $\boldsymbol{bins}$
normalize the bins using (3)
$i = 1$ #indexing starts at 0
$numBinsM = 1$ #num of bins merged with current bin
while $i < |\boldsymbol{bins}|$ do
    $avgSizeOfMerged = \frac{\boldsymbol{bins}^n_{i-1}}{numBinsM}$
    $maxBinSize = max(avgSizeOfMerged, \boldsymbol{bins}^n_i)$
    $minBinSize = min(avgSizeOfMerged, \boldsymbol{bins}^n_i)$
    $maxMinSize = maxBinSize + minBinSize$
    $perc = \frac{minBinSize}{maxMinSize}$
    $\mathbf{CI} = confInterval(perc, maxMinSize, 0.05)$
    if $50\% \in \mathbf{CI}$ then
        $\boldsymbol{bins}^{right}_{i-1} = \boldsymbol{bins}^{right}_i$
        $\boldsymbol{bins}^n_{i-1} = \boldsymbol{bins}^n_{i-1} + \boldsymbol{bins}^n_i$
        $\boldsymbol{bins} = \boldsymbol{bins} - \boldsymbol{bins}_i$
        $numBinsM = numBinsM + 1$
    else
        $\boldsymbol{bins}^n_{i-1} = avgSizeOfMerged$
        $numBinsM = 1$
    end if
    $i = i + 1$
end while
return $\boldsymbol{bins}$
```

instances, and stricter with bins of high density. This will enable the identification of intervals in which the densities have a significant difference. This can be tested with a two-tailed test of a population proportion, in which the population is the number of instances in the two bins and the hypothesized value of the true population proportion between the two bins is 50%. Essentially the proportion test decides that two bins are similar if the calculated confidence interval contains 50%.

The steps of *mergeSimilarDensityBins* are shown in Algorithm 5. Initially, two consecutive bins are selected and the bin with the largest and smallest number of instances out of the two is identified (steps 6-7). Then the percentage of the instances of the smallest bin to the sum of the instances of the two is calculated in step 9. Using the calculated percentage and the number of instances of the two bins, the confidence interval at a 5% significance level is calculated (step 10). If 50% is within the calculated confidence interval, the two bins get merged. Initially the right interval of the left bin is set equal to the right interval of the right bin (step 12), and the size of the right bin is added to the size of the left bin (step 13). The right bin is then removed from the ***bins*** vector (step 14) and the counter of the number of bins contained in the current bin is incremented by one (step 15). Thus, once two or more bins get merged, their new size is the mean number of instances of the bins used to create it (step 5).

Table 8: Confidence Intervals as the Number of Instances Increases

$\boldsymbol{bins}_1^n$	$\boldsymbol{bins}_2^n$	$\boldsymbol{bins}_1^n/(\boldsymbol{bins}_1^n + \boldsymbol{bins}_2^n)$	**CI**
10	11	47.6%	$26.4\% - 69.7\%$
100	110	47.6%	$40.7\% - 54.6\%$
1000	1100	47.6%	$45.5\% - 49.8\%$
10000	11000	47.6%	$46.9\% - 48.3\%$

The range of a confidence interval is affected by the population size, since the larger the population the more confident we are on a percentage. In this case, the population size is the number of instances that belong to the two bins that DBAD is trying to merge. Specifically, the range of the confidence interval gets narrower as the number of instances increases and hence the criterion becomes stricter. Table 8 represents the number of instances of two bins ($\boldsymbol{bins}_1^n, \boldsymbol{bins}_2^n$), the percentage of bin1 ($\boldsymbol{bins}_1^n/(\boldsymbol{bins}_1^n + \boldsymbol{bins}_2^n)$) and the calculated confidence interval (CI) at a 5% confidence level. As can be seen, for the same percentage, only in the first two cases the 50% criterion is met, but DBAD should merge the two bins in all 4 cases, since 47.6% is very close to 50%.

Hence, this approach behaves as needed on low density bins but can be too strict as the bins density increases. For such bins, the criterion should be strict but should still be able to consider them as similar if their ratio is above a certain threshold. The narrowest confidence interval will be the one testing if the largest bin should be considered similar to one of its neighboring bins (if they indeed have a similar number of instances). Thus, to control the strictness of the metric, the sample size of the bins needs to be normalized. To accomplish this, in case there are more than α instances in the largest bin ($\boldsymbol{bins}_{max}$), the number of instances in each bin gets normalized using (3), so that $\boldsymbol{bins}_{max}$ ends up having α instances.

$$\boldsymbol{bins}_i^n \leftarrow \lceil \boldsymbol{bins}_i^n * \frac{\alpha}{\boldsymbol{bins}_{max}^n} \rceil \tag{3}$$

To select the α value, it is needed to define one more threshold (t). This threshold will control when a bin compared to the largest one should be considered having a similar density. For DBAD it was empirically selected to consider a bin that is consecutive to the largest one as similar, if it has at least 75% of its instances ($t = 75\%$).

To find the α value for which t is satisfied, the following process is performed. For each α value (size of the largest bin) that will be tested, the minimum size of a consecutive bin that passes the proportion test is identified and its ratio to α is calculated. Then the α value for which the ratio is equal to t is selected. In Figure 6, the points indicate the ratio of the smallest bin size that will

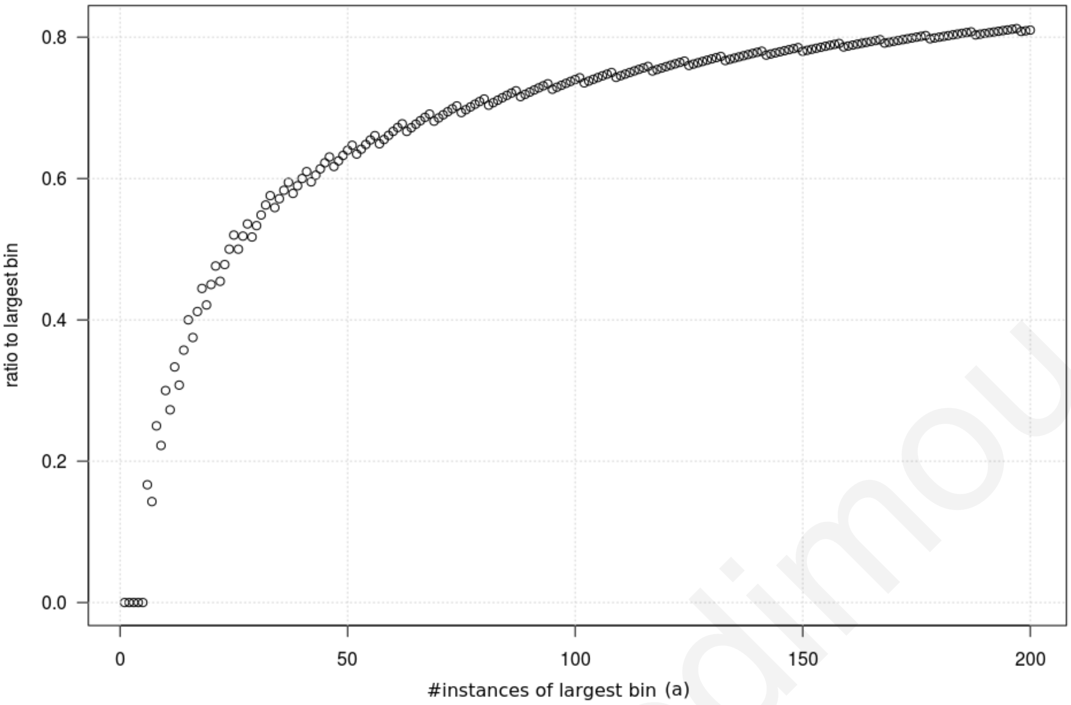


Figure 6: The *x* axis represents the number of instances of the largest bin (α) and the *y* axis the ratio of the number of instances of a bin to the largest. The points indicate the ratio of the smallest bin size that will be considered having a similar density to an α sized bin.

be considered having a similar density to an α sized bin. As expected, with a larger number of instances (α) the more the ratio increases, since the calculated confidence interval gets narrower and its lower bound increases. To obtain the targeted $t = 75\%$, for DBAD α is set to 112, thus the largest bin will have at most 112 instances after the normalization.

In the next step of DBAD (Algorithm 3), the intervals of the remaining bins are merged with the intervals identified for other classes of the feature. Once this is done for all of the classes, *removeEmptyBins* is executed on them so as to remove any possible empty bins that we might have after the merging of the intervals of each class. These are the final intervals that will be used for discretizing the feature.

4.2.2 Parallelization

The algorithm of DBAD, can be easily parallelized and executed by multiple resources, since each feature can be discretized independently. This is also known as an embarassingly parallel problem, that can achieve high-levels of performance. A hybrid parallel versio of DBAD that runs on both distributed and shared memory systems using the Message Passing Interface (MPI) [69] and OpenMP [70] has been implemented in C++. The algorithm splits the execution to multiple MPI processes, where each node takes one equal portion of the dataset. Each MPI process is scheduled on a separate node in the system, where the DBAD discretization is further parallelized using OpenMP threads (see Figure 7).

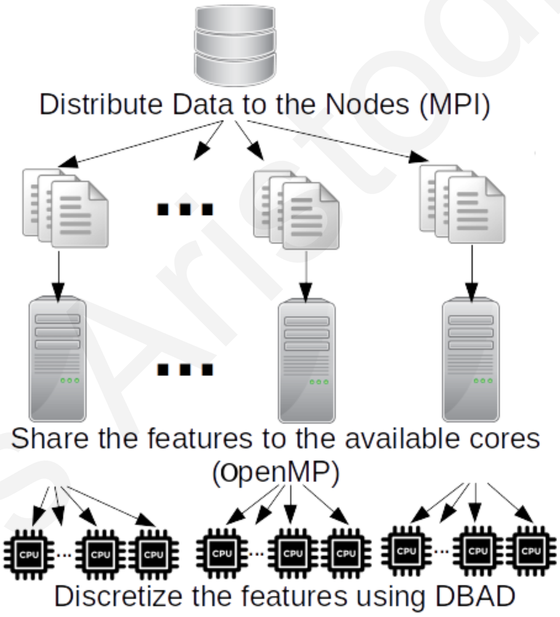

Distribute Data to the Nodes (MPI)
...
...
Share the features to the available cores (OpenMP)
CPU
CPU
CPU
CPU
CPU
CPU
CPU
CPU
CPU
Discretize the features using DBAD

Figure 7: Parallelization of DBAD.

The algorithm gets as input multiple splitted files of the initial input file, so as to allow processing big data that cannot fit in the RAM of a single node and at the same time reduce the communication overhead between parallel nodes. Initially MPI is used to distribute the input files evenly to the available nodes. Then each node reads one file at a time and splits its features to the available cores it has and each core executes DBAD on its provided features. When the cores finish with the discretization of a file, the node returns the discretized version of the file, where the master node will merge all results to produce the final output.

4.2.3 Datasets

A summary of the datasets used in the evaluation is given in Table 9. All of the datasets are from the UCI Machine Learning Repository [40], except from the two synthetic datasets that are used for testing the speedup of the parallel version of DBAD.

Table 9: Summary of the datasets investigated

Dataset	#Instances	#Features	Instances per Class
Heart	270	13 (7)	150 / 120
SPECTF	267	44 (0)	55 / 212
Cardiotocography	2126	35 (11)	1655 / 295 / 176
Diabetic retinopathy	1151	19 (3)	560 / 611
Haberman	306	3 (0)	225 / 81
Saheart	462	9 (1)	302 / 160
WisconsinBC	569	30 (0)	357 / 212
Pima	768	8 (0)	500 / 268
ILPD	583	10 (1)	416 / 167
Mammographic	830	5 (4)	427 / 43
Arcene	200	9961 (0)	112 / 88
Synthetic1	10000	1000000 (0)	5000 / 5000
Synthetic2	1000000	10000 (0)	500000 / 500000

Number of categorical features given in parentheses

The datasets from the UCI repository, cover a wide range of datasets of the life sciences domain. They were selected so as to have different ranges of number of instances and features, and have a variety of class balances. These datasets had a two step preprocessing procedure. The first step was to replace any missing values using (4), which basically replaces the missing values of a variable with its minimum value minus one.

$$\mathbf{x}_{\emptyset} \leftarrow min(\mathbf{x}) - 1 \tag{4}$$

The second was to remove any single valued features since they did not contain any information that could help with the classification.

The two synthetic datasets, were created using different continuous distributions (uniform, Gaussian and beta) on each feature, so as to cover some of the common distributions a dataset might have. The Synthetic1 dataset has a large number of features, whereas Synthetic2 a large number of instances, so as to see if the algorithm is affected by the size of each dimension.

4.2.4 Evaluation Methodology

To evaluate the effectiveness of DBAD at producing meaningful intervals, a 10 fold stratified Cross Validation (CV) methodology was followed. Initially the training folds were used by a discretizer so as to create the intervals for each feature. Using these intervals, the training and testing folds were discretized. Then a classifier was trained on the training folds and its accuracy and Cohen's kappa value was measured on the testing fold. Specifically, the Support Vector Machine (SVM) [44], the Random Forest [23] and Naive Bayes were used for the classification task.

Additionally, the inconsistency rate of the discretizer was calculated on the testing fold. The inconsistency rate is calculated on each feature and it measures the percentage of instances that will be unavoidably misclassified if a single discretized feature is used in a classification. This is due to the fact that after the discretization of a feature, some instances will end up having the same value but their class will differ.

The evaluation was performed among DBAD and three more discretizers, which were documented to perform well in the literature [50]. The selected discretization algorithms used, were CAIM and ChiMerge which offer excellent performances in different types of classifiers, and MDLP which provides a good trade-off between the number of produced intervals and accuracy

[50]. All algorithms used the same folds for training and testing during the CV so as to obtain comparable results.

To test if the intervals created by each discretization algorithm were superior or inferior to DBAD's intervals, Wilcoxons' signed-rank test was used on the evaluation measures mentioned above and additionally on the number of intervals produced by each algorithm and its execution time. The difference between two algorithms is considered statistically significant, if the obtained p-value from the test is less than or equal to 0.05.

The aforementioned evaluation was performed in R. Specifically, from the *e1071* package [71] the SVM and Naive Bayes were used, from the *randomForest* package [48] Random Forest was used, whereas for the discretization part, the *discretization* package [72] was used for the three discretizers. All of the aforementioned algorithms were used with their default parameters. The version of DBAD used for the aformentioned evaluation was implemented in R so as to have comparable results with the other discretization methods regarding their execution time which were also in R.

For testing the speedup of the parallel version of DBAD, Synthetic1 and Synthetic2 were split to 10 files each. Specifically, Synthetic1 was split to 10 files with 10,000 instances and 100,000 features each, whereas Synthetic2 was split to 10 files with 1,000,000 instances and 1,000 features each. The experiments were run at the Cyprus Institute HPC facility (CY-TERA), on nodes with 48 GB of RAM and the Intel Westmere X5650 hexa-core CPUs. The speedup, was tested using a different number of nodes (1-10) and threads (1,2,4,6).

4.3 Results

In this section the evaluation results of DBAD are presented. Initially results regarding the number of intervals created by each algorithm and the inconsistency rate are presented. Then metrics regarding the performance of the classifiers are shown, which can test the goodness of the intervals and the information lost. Additionally the execution times of each algorithm are shown and finally the obtained speedup with the parallel version of DBAD is provided. To show if DBAD performed better or worse than another algorithm, the result of the statistical analysis is shown with a sign next to the accuracy. A positive sign indicates that DBAD performed better, whereas a negative sign shows that DBAD had worse performance.

Table 10: Average Number of Intervals

Dataset	**DBAD**	**MDLP**	**CAIM**	**ChiMerge**
Heart	5.0	1.7 (-)	2.0 (-)	4.5 (-)
SPECTF	6.7	1.5 (-)	2.0 (-)	4.0 (-)
Cardiotocography	8.7	3.0 (-)	2.7 (-)	14.2 (+)
Diabetic retinopathy	7.5	1.8 (-)	2.0 (-)	46.8 (+)
Haberman	5.0	1.3 (-)	2.0 (-)	3.1 (-)
Saheart	7.4	1.7 (-)	2.0 (-)	17.7 (+)
WisconsinBC	10.1	3.0 (-)	2.0 (-)	44.2 (+)
Pima	9.1	2.1 (-)	2.0 (-)	13.5 (+)
ILPD	7.1	1.6 (-)	2.0 (-)	10.9 (+)
Mammographic	7.2	2.3 (-)	2.0 (-)	3.7 (-)
Arcene	4.0	1.3 (-)	2.0 (-)	7.4 (+)
average	**7.1**	**1.9**	**2.1**	**15.5**

(+) DBAD was statistically better, (-) DBAD was statistically worse

For comparing the produced number of intervals by each discretizer, Table 10 provides the average number of intervals created by each algorithm on the training data. In most of the datasets, the smallest number of intervals was created by MDLP and then by CAIM. DBAD is third and created 7 intervals on average on each dataset, whereas ChiMerge had the largest number of intervals in most of the tests.

Table 11: Average Inconsistency Rates

Dataset	DBAD	NoDiscr	MDLP	CAIM	ChiMerge
Heart	32.3	23.6 (-)	34.5 (+)	33.3 (+)	31.9
SPECTF	19.3	11.5 (-)	20.6 (+)	20.4 (+)	20.0 (+)
Cardiotocography	20.1	17.5 (-)	20.2 (+)	20.2 (+)	19.5 (-)
Diabetic retinopathy	41.2	18.7 (-)	45.0 (+)	43.7 (+)	33.2 (-)
Haberman	25.0	16.0 (-)	25.7	25.5	25.4
Saheart	30.4	10.6 (-)	33.8 (+)	32.3 (+)	27.0 (-)
WisconsinBC	20.4	0.4 (-)	22.9 (+)	22.9 (+)	14.4 (-)
Pima	30.0	16.0 (-)	33.1 (+)	32.4 (+)	29.2
ILPD	27.3	16.7 (-)	28.6 (+)	28.5 (+)	26.7 (-)
Mammographic	27.3	24.8 (-)	28.3 (+)	27.9 (+)	27.6
Arcene	38.7	19.4 (-)	42.2 (+)	39.6 (+)	33.4 (-)
average	**28.4**	**15.9**	**30.4**	**29.7**	**26.2**

(+) DBAD was statistically better, (-) DBAD was statistically worse

The inconsistency rate of the testing folds is presented in Table 11. The lowest inconsistency rate was obtained with the original data (NoDiscr), which is essentially the minimum error that can be obtained by a discretization algorithm. The discretization algorithm with the lowest inconsistency rate is ChiMerge and then DBAD follows. MDLP had the worst performance from the discretization algorithms regarding this measure.

Table 12: Classification Accuracy

Dataset	DBAD	NoDiscr	MDLP	CAIM	ChiMerge
Support Vector Machine results					
Heart	83.7	83.7	84.4	84.1	80.0
SPECTF	78.7	79.0	78.3	79.2	80.6
Cardiotocography	98.8	98.7	99.1	98.8	98.9
Diabetic retinopathy	68.2	69.5	62.3 (+)	63.3 (+)	53.2 (+)
Haberman	72.9	74.8	71.9	74.5	73.2
Saheart	72.1	71.6	70.3	71.2	71.6
WisconsinBC	97.9	97.2	97.0	94.9 (+)	97.2
Pima	75.8	75.5	76.3	73.1	75.4
ILPD	71.3	71.2	70.5	68.2	70.8
Mammographic	82.4	82.9	84.0	83.7	83.1
Arcene	82.4	81.9	66.4 (+)	79.3	70.9 (+)
Random Forest results					
Heart	85.6	84.8	85.2	84.8	84.4
SPECTF	81.4	82.5	79.8	81.4	81.8
Cardiotocography	98.9	98.8	99.0	98.8	98.8
Diabetic retinopathy	66.9	68.4	62.6 (+)	63.9 (+)	65.5
Haberman	72.6	74.5	72.6	73.5	74.5
Saheart	72.7	68.4 (+)	70.1	73.0	69.0 (+)
WisconsinBC	96.0	96.5	96.7	95.4	96.8
Pima	77.1	76.8	75.8	75.1	75.4
ILPD	69.3	70.0	69.8	67.9	70.1
Mammographic	81.9	80.6	83.6	84.2 (-)	82.8
Arcene	80.5	82.9	86.0	81.4	82.9
Naive Bayes results					
Heart	83.3	78.9	84.4	84.1	81.1
SPECTF	77.6	77.6	77.2	78.3	81.0
Cardiotocography	95.5	95.8	96.6 (-)	96.2	96.4
Diabetic retinopathy	64.4	61.4	62.6	61.8	61.9
Haberman	74.5	71.6	71.9	75.8	73.2
Saheart	68.2	67.5	69.7	72.3	66.3
WisconsinBC	93.8	73.6 (+)	94.2	94.4	93.8
Pima	74.8	66.9 (+)	76.6	72.5	73.7
ILPD	66.9	66.0	67.7	66.4	66.4
Mammographic	83.0	82.8	82.4	81.6 (+)	83.0
Arcene	70.9	67.4	66.4 (+)	66.9	73.0
average	**79.5**	**78.2**	**78.5**	**78.8**	**78.4**

(+) DBAD was statistically better, (-) DBAD was statistically worse

Table 12 presents the obtained accuracies with the discretized data of each algorithm using the three classifiers. The NoDiscr column shows the accuracies obtained with the non-discretized data. As can be seen, DBAD had a similar performance with the other algorithms for the majority of the tests (87%) and performed better (won) 11 times (11%), while it performed worse (lost) in only 2 tests (2%). Additionally, the algorithm had a similar or better performance than using the non-discretized data and on average had the highest accuracy from all of the tested approaches.

Table 13: Sensitivity Specificity and Precision

	DBAD	**NoDiscr**	**MDLP**	**CAIM**	**ChiMerge**	
Sensitivity						
average	75.3	71.7	73.9	74.5	73.0	**total**
SVM		(2,7,1)	(3,6,1)	(2,8,0)	(1,8,1)	(8,29,3)
RF		(0,10,0)	(1,7,2)	(0,9,1)	(0,10,0)	(2,36,2)
NB		(5,4,1)	(1,8,1)	(3,5,2)	(2,7,1)	(11,24,5)
total		(7,21,2)	(5,21,4)	(6,22,2)	(3,25,2)	(21,89,10)
Specificity						
average	70.3	68.4	68.7	69.6	69.6	**total**
SVM		(1,6,3)	(1,9,0)	(1,7,2)	(1,7,2)	(4,29,7)
RF		(1,8,1)	(4,5,1)	(3,6,1)	(1,9,0)	(9,28,3)
NB		(2,7,1)	(1,9,0)	(1,8,1)	(1,7,2)	(5,31,4)
total		(4,21,5)	(6,23,1)	(5,21,4)	(3,23,4)	(18,88,14)
Precision						
average	75.2	70.3	69.3	73.2	72.4	**total**
SVM		(1,8,1)	(4,6,0)	(2,7,1)	(2,7,1)	(9,28,3)
RF		(1,9,0)	(4,6,0)	(3,7,0)	(1,9,0)	(9,31,0)
NB		(2,8,0)	(1,9,0)	(1,9,0)	(0,10,0)	(4,36,0)
total		(4,25,1)	(9,21,1)	(6,23,1)	(3,26,1)	(22,95,3)

(wins, ties, losses)

Table 13 shows the performance of the algorithms based on their sensitivity, specificity and precision. The cardiotocography dataset was not included in this analysis since it has three classes. Regarding the sensitivity, DBAD had a similar number of wins and losses against each of the other discretization algorithms, except CAIM from which it had more wins, whereas in specificity it had more wins against MDLP. DBAD, performed better more times than the other discretizers in precision. Overall, DBAD performed better than the other algorithms and the non discretized data in 61 tests and worse in 27, and on average had higher values in all three measures.

Table 14: Cohen's Kappa Values

Dataset	DBAD	NoDiscr	MDLP	CAIM	ChiMerge
Support Vector Machine results					
Heart	0.67	0.67	0.68	0.67	0.59
SPECTF	0.19	0.01 (+)	0.00 (+)	0.25	0.29 (-)
Cardiotocography	0.97	0.97	0.97	0.97	0.97
Diabetic retinopathy	0.37	0.39	0.26 (+)	0.27 (+)	0.00 (+)
Haberman	0.18	0.18	0.12	0.25	0.20
Saheart	0.33	0.32	0.30	0.34	0.32
WisconsinBC	0.95	0.94	0.94	0.89 (+)	0.94
Pima	0.44	0.44	0.44	0.37	0.43
ILPD	0.09	0.00 (+)	-0.02 (+)	0.09	0.13
Mammographic	0.65	0.66	0.68	0.67	0.66
Arcene	0.65	0.63	0.34 (+)	0.58	0.37 (+)
Random Forest results					
Heart	0.71	0.69	0.70	0.69	0.68
SPECTF	0.28	0.29	0.34	0.34	0.32
Cardiotocography	0.97	0.97	0.97	0.97	0.97
Diabetic retinopathy	0.34	0.37	0.27 (+)	0.28 (+)	0.31
Haberman	0.14	0.20	0.16	0.11	0.18
Saheart	0.37	0.25 (+)	0.29	0.37	0.28 (+)
WisconsinBC	0.91	0.92	0.93	0.90	0.93
Pima	0.47	0.48	0.43	0.41 (+)	0.44
ILPD	0.20	0.18	0.02 (+)	0.08	0.22
Mammographic	0.64	0.61	0.67	0.68 (-)	0.66
Arcene	0.66	0.65	0.71	0.61	0.65
Naive Bayes results					
Heart	0.66	0.57	0.68	0.68	0.62
SPECTF	0.42	0.24	0.45	0.47	0.44
Cardiotocography	0.88	0.89	0.91 (-)	0.90	0.90
Diabetic retinopathy	0.29	0.24	0.27	0.22 (+)	0.24
Haberman	0.21	0.14	0.12	0.28	0.20
Saheart	0.31	0.26	0.34	0.38	0.26
WisconsinBC	0.87	0.41 (+)	0.88	0.88	0.87
Pima	0.45	0.27 (+)	0.48	0.37 (+)	0.42
ILPD	0.31	0.23	0.31	0.27	0.25
Mammographic	0.66	0.66	0.65	0.63	0.66
Arcene	0.42	0.36	0.34	0.35	0.45
average	**0.51**	**0.46**	**0.47**	**0.49**	**0.48**

(+) DBAD was statistically better, (-) DBAD was statistically worse

The obtained Cohen's Kappa values are shown in Table 14. Similar to the accuracy results of Table 12, DBAD had a similar performance with the other algorithms in most of the performed tests. Overall, it performed better than the other algorithms in more tests (15) than it did not (3). Most of its wins were with the SVM classifier, as was the case with the accuracy tests.

Table 15: Average Execution Time in Seconds

Dataset	DBAD	MDLP	CAIM	ChiMerge
Heart	0.0	0.1 (+)	0.4 (+)	1.4 (+)
SPECTF	0.3	0.7 (+)	2.1 (+)	3.8 (+)
Cardiotocography	0.2	2.7 (+)	18.7 (+)	100.1 (+)
Diabetic retinopathy	0.1	3.6 (+)	28.8 (+)	453.4 (+)
Haberman	0.0	0.0 (+)	0.2 (+)	0.2 (+)
Saheart	0.1	0.8 (+)	3.3 (+)	29.3 (+)
WisconsinBC	0.2	10.2 (+)	31.1 (+)	408.2 (+)
Pima	0.1	0.8 (+)	3.1 (+)	21.8 (+)
ILPD	0.1	0.6 (+)	2.2 (+)	8.9 (+)
Mammographic	0.0	0.1 (+)	0.2 (+)	0.3 (+)
Arcene	68.9	301 (+)	632 (+)	2,199 (+)
average	**6.3**	**29.1**	**65.7**	**293.4**

(+) DBAD was statistically better, (-) DBAD was statistically worse

The average execution time of each discretization algorithm on the training folds is provided in Table 15. DBAD has the lowest computational complexity from the rest of the algorithms, which explains the results shown in Table 15. DBAD is the fastest and in some tests it is at least an order of magnitude faster than the rest. The difference is more obvious in the Arcene dataset, which has the largest number of features from all the datasets. In this dataset, DBAD is approximately 4 times faster than MDLP, 9 times faster than CAIM and 32 times faster than ChiMerge.

Table 16: Obtained Speedup

#Nodes #Threads	Dataset	1	2	3	4	5	10
1	**Synthetic1**	1.0	2.0	2.5	3.3	4.9	9.9
	Synthetic2	1.0	2.0	2.5	3.3	4.9	9.6
2	**Synthetic1**	1.7	3.5	4.3	5.7	8.6	17.2
	Synthetic2	1.8	3.4	4.3	5.7	8.6	17.1
4	**Synthetic1**	2.7	5.4	6.6	9.1	13.5	26.3
	Synthetic2	2.8	5.6	7.0	9.4	14.1	28.0
6	**Synthetic1**	3.3	6.7	8.4	11.1	16.4	33.3
	Synthetic2	3.8	7.0	8.9	11.8	17.8	35.3

The results using 6 to 9 nodes are omitted since they had a similar speedup with 5 nodes.

Table 16, provides the obtained speedup of the parallel version of DBAD. The time needed to discretize the Synthetic1 dataset with a single node and a single thread was 233 minutes (approximately 3.9 hours), whereas Synthetic2 needed 217 minutes (7% faster). The first thing worth mentioning, is that the speedup is similar in both synthetic datasets, which indicates that the parallel execution scales well with the size of the dataset. The second thing we notice is that on the single thread executions, there is a linear speedup when the number of splits (10) is divisible by the number of nodes. The last thing of interest is that if we consider the change of speedup when we keep the number of nodes constant and increase the number of threads, is that the obtained speedup is not linear. In this scenario, performance scalability is limited by the size of the cache in the processor. The large size of the input file doesn't fit in the processors cache, thus requires accessing DRAM.

4.4 Discussion

An important feature of a discretization algorithm, is to be able to produce a small number of intervals. Even though MDLP and CAIM performed better on this task, DBAD produced a reasonable number of intervals, whereas ChiMerge produced the largest number of intervals. DBAD might produce a large number of intervals in the case of datasets with many classes with different distributions. This is due to the fact that the intervals are created for each class separately and if the densities are in different ranges for each class, we will end up with much more intervals. On the other hand this might not affect a classifier since it could further help with the separation of the classes.

The results regarding the inconsistency rates seem to be associated with the number of intervals produced by the algorithms. As was also found by [50], the more intervals an algorithm creates, the lower its inconsistency rate is. The performance on this metric does not seem to be correlated with the performance of the classifiers. If that was the case, then the undiscretized data would have produced better results than the discretized.

The results regarding the obtained accuracies by the discretized data of each algorithm, indicate that DBAD has a similar or better performance than the other algorithms in the majority of the tests and in only 2% of the performed tests its performance was worse. Additionally, when using the SVM classifier, it did not perform worse than any of the other methods. A positive outcome of this analysis is that DBAD's discretization did not significantly reduce the information of the datasets, since it performed equally or better than the original datasets in all of the tests. The results on specificity indicate that it had a similar performance with the other methods on this measure, while it was able to perform better than the other algorithms more times than they did when considering sensitivity and precision. Cohen's kappa values also had more tests in which DBAD had a better performance than the other discretization algorithms (15 wins and 3 losses). Since this measure takes into consideration random hits [73], and many of the tested datasets were also imbalanced, it is a more suitable measure for comparing the efficiency of the discretization algorithms. On average, DBAD had the highest scores in all of the aforementioned measures and since the statistical analysis indicates that it performs similarly or better than the other algorithms in the majority of the tests, there is a good indication that using the densities of each class is a promising approach for data discretization.

The proposed algorithm has a clear advantage over the others, regarding its computational complexity. A dataset can be discretized by DBAD in $O(m*n)$, where m is the number of features and n is the number of instances. Specifically, for a single feature, the binning and number of instances in each bin can be calculated with a single pass from the feature's data ($O(n)$), whereas the removal of empty bins and the merging of similar density bins is $O(\log(n))$, since this is approximately the number of bins that are initially created using Sturge's rule. Algorithms like

CAIM, MDLP and ChiMerge, have a computational complexity of at least $O(n * log(n))$ for each feature, since this is the complexity for sorting the data. Then the Top-Down approaches have an additional cost of $O(k * n)$, where k is the number of intervals created. For CAIM and MDLP this value is close to the number of classes and since this is much smaller than the number of instances their complexity is $O(m * n * log(n))$. Due to the fact that ChiMerge, merges two consecutive points/bins at a time, this means that it has an additional complexity of $O((n - k) * n)$, and hence its complexity for discretizing a dataset is $O(m * n^2)$. Based on this analysis, it is clear that DBAD has a low computational complexity, which explains its small execution times on the performed experiments.

In addition to its low complexity, DBAD is an embarrassingly parallel algorithm, since each feature can be discretized separately, which makes the algorithm suitable for big data. When using multiple threads, after a point the speedup is not linear because the overhead of reading the dataset in memory and writing the discretized version of the dataset is larger or equal to the time for discretizing the data. This is because the reading and writing tasks are sequential and cannot be parallelized and once the parallel task of the discretization becomes faster than the time needed to read and write the data, no additional speedup can be obtained. On the other hand, the speedup analysis, shows that it can fully exploit the use of distributed systems since the reading and writing of each file can be done in parallel in each node. But to get the optimal speedup, one should split the initial data to a number of files that is divisible with the number of nodes that will be used. This will allow an optimal load balancing to each node.

In general, DBAD is a good choice for discretizing big data. A limitation of the algorithm is that it will under-perform if the data of each class to be discretized are from the same uniform distribution. In such cases, since the density is the same in the entire range of values for all classes, DBAD will not be able to identify any intervals and will return a single valued feature. On the other hand, if data are from different distributions, since DBAD uses the density change and not a direct measure that checks for class separability, it can have an advantage over such methods in cases that have interacting features with low main effects. In such cases, a single feature does not have enough information for separating the different classes, unless it is considered in combination with other features, and hence a method that uses a measure based on class separability will return a single valued feature. DBAD, on the other hand, would still be able to split such features based on their density differences, and those splits could allow the identification of interacting features.

4.5 Conclusions

A new supervised discretization algorithm has been proposed, based on how the density of the values of a feature changes for each class. DBAD is a univariate discretizer and hence does not consider any possible feature interactions when selecting the best intervals for each feature, but results indicate that it does not significantly reduce the information contained in a dataset. It is a static discretizer since it is independent of the classifier that will be used at the processing step, which allows it to produce more generic intervals that can perform well with different classifiers. It initially creates all of the bins simultaneously and then follows a bottom-up approach for merging consecutive bins with similar densities, based on the confidence interval of their percentage.

In the performed evaluation, DBAD produced an acceptable number of intervals and had comparable results with the other discretizers. An advantage it has, over the other methods, is its low computational complexity, which makes it appropriate for big data. Another reason DBAD can be easily applied on big data, is because it is an embarrassingly parallel algorithm, since each feature can be discretized independently, and our experiments have shown that it can obtain a linear speedup.

The results of the evaluation, indicate that density based discretization is a promising approach and needs to be further investigated. One possible direction is to test different methods for the initial binning of the data, since the current approach is based only on the number of instances and hence the true distribution of the data might be misrepresented. Another possible direction could be on having an additional step for further reducing the number of the final intervals of each feature based on other metrics.

Chapter 5

Clustering Subjects in Genetic Studies with Self Organizing Maps [4]

5.1 Introduction

In common complex diseases, multi-loci interactions are more important than the main effect of any single SNP. Single locus association studies in such diseases may not replicate their results across multiple samples, due to the effect of epistasis and other phenomena [6]. In this chapter, Self Organizing Maps (SOMs) are investigated for clustering Genome Wide Association (GWA) data for multi-loci association testing.

Traditional genetic analyses focus on single locus associations and not on multi-loci associations. Several machine learning techniques have been applied for multi-loci association testing [74, 75, 76]. Most of them are using Neural Networks (NNs), Support Vector Machines (SVMs), Random Forests (RFs), Multifactor Dimensionality Reduction (MDR) and variations of these techniques and they will be introduced.

Ritchie *et al* applied Multifactor Dimensionality Reduction (MDR) to a sporadic breast cancer data set [77], whereas variations of MDR were also introduced for multi-loci associations [78, 79, 80]. MDR and its variations had good results when they were used on a small number of SNPs but they cannot be directly used on a large number of SNPs, due to the exhaustive search it performs for identifying n-SNP associations [76].

NNs have also been applied in such studies [81], but their results are affected by the architecture of the network. Since it is computationally intractable to perform an exhaustive search for selecting the appropriate architecture, techniques based on Genetic Programming Neural Networks (GPNN) [82] and Grammatical Evolution Neural Networks (GENN) [83] were proposed with promising results. When these two optimisation methods were compared, GENN outperformed GPNN [83].

Wadell *et al* [84] used SVM to test their hypothesis that different SNP patterns exist among patients with Multiple Myeloma diagnosed at a young age and patients diagnosed after the age of 70. They obtained an accuracy of 71% in classifying these two patient classes giving them evidence that their initial hypothesis was correct. In [85] an SVM approach was proposed for gene-gene interaction with comparable results with MDR. The approach handled unbalanced data better than MDR and was less susceptible to overfitting, but it was computationally expensive [85]. A disadvantage of SVMs, is that they do not cope well with missing data [85].

RFs have also been used for multi-loci association testing [86, 87]. A variation of RF is SNPInterForest [88], which copes with some limitations of RFs such as the fact that they may ignore SNPs with low marginal effects and the difficulty of extracting the interactions patterns. SNPInterForest was applied on 10,000 SNPs and identified two novel interactions but the analysis was computationally demanding. Another variation of RFs is Random Jungle (RJ) [89], which is an efficient method for analysing GWA data. The method was applied on 275,153 SNPs revealing new interactions and validating findings of recent studies. As with RFs, RJ's findings were affected when SNPs only had weak main effecs [76].

SOMs [90] were proposed by Tuevo Kohonen and have been widely used for clustering data in several scientific areas. To the authors knowledge, they have not been used for identifying multi-loci associations, but they have been used on biological data analyses [91, 92, 93, 94]. Classical SOMs were intended for use with numerical data, but SNPs are nominal categorical data, hence a SOM for categorical data needs to be used for such data. NCSOM [95] is an algorithm that was proposed for handling numerical and categorical variables with promising results, and its update rule for nominal categorical variables is also used in this work.

In this study, the NCSOM algorithm was tested on GWA data, and specifically on subjects with and without Multiple Sclerosis (MS). Initially a subset of the SNPs was selected and then the algorithm was trained with 10-fold cross validation for selecting the appropriate map size. Once the map size was defined, the algorithm was applied on the selected SNPs and its clustering results were evaluated.

The structure of the chapter is as follows. Section 5.2 provides the methodology followed for the experiments carried out and in Section 5.3 the results of the experiments are presented. In Section 5.4 the results are discussed and finally in Section 5.5 the concluding remarks are provided.

5.2 Methodology

5.2.1 Dataset

The data used are from Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene)[96]. The dataset has 1,618 people with MS (cases) and 3,413 people without MS (controls). The genotyping platform used was made by Illumina 300k model. The distribution of the subjects in the dataset is shown in the first three columns in Table 17.

Table 17: ANZgene Subjects Distribution

	Males	Females	Total	Training	Testing
Cases	445	1173	1618	1456	162
Controls	757	1231	1988	1789	199
Total	1202	2404	3606	3245	361

Before using the data for training SOM, they were encoded using the encoding schema shown in Table 18. An allele value of ”a” represents the minor allele of a SNP and ”A” represents the major allele of that SNP. With this encoding, if a subject has homozygous minor allele in a SNP and another subject has homozygous major allele on that SNP the distance between the two subjects for that SNP will be two. In the case that one of the subjects has a heterozygous allele and the other has a homozygous allele the distance will be one. The missing alleles were encoded with a -1 so that SOM can identify them, since it handles missing data differently.

Table 18: Data Encoding

Allelle 1	Allele 2	Encoding Value
a	a	0 1
a A	A a	1 1
A	A	1 0
Missing	Missing	-1 -1

5.2.2 Feature Selection

The dataset consists of approximately 300,000 SNPs per subject, but this work focuses on SNPs in the HLA region, where previous studies have identified associations among the region and MS [97, 98, 99, 100]. A subset of SNPs in the regions was selected using a two SNP interaction algorithm [101, 102]. Specifically, the SNP pairs that were found to be associated with the disease by the two SNP interaction algorithm were selected as input for the training of SOM. This analysis resulted in a total of 37 SNPs. The advantage of using a two SNP interaction algorithm for feature selection with SOM, is that we can test for associations among the selected SNPs. For example if the two SNP interaction algorithm returned that SNP A and SNP B were associated and that SNP A and SNP C were associated, SOM would also cluster the association of SNP A, B and C with the disease.

5.2.3 Categorical SOM for clustering SNPs

A batch categorical SOM was used, with the update rule for nominal categorical data from [95] with some modifications for handling missing data. Specifically, missing data were ”ignored”, since they were always considered as a match when compared with any allelic value. SOM consisted of an input layer which selected an input vector at a time as the input of the network, and an output layer that had the neurons that represented the final clusters. The input layer was an N dimensional vector **x**, where N was the number of features of the P input vectors. The output layer consisted of the neurons mapped in a two dimensional map. The number of neurons and the size of the map was predefined before training. Each neuron consisted of an N dimensional vector **w** called the weights of the cluster and the ith weight of each cluster corresponded to the ith feature of the input vector. The set **a** $= \{0, 1, -1\}$ represents the possible categorical values of each input vector feature and $\mathbf{a}_r$ represents the categorical value with index r, where $r = 1, 2, 3$.

Algorithm 6 Categorical SOM

```
  Initialize the weights (w) with random values $\{0,1\}$
  $t = 1$ (current epoch)
Require: $radius$ (neighbourhood radius), $T$ (final epoch)
  while $t \neq T$ do
    for each input vector $\mathbf{x}^j$ do
      find its BMU using (5)
    end for
    update the weights of each neuron using (8)
    $t = t + 1$
    reduce $radius$
  end while
```

The training of the network is shown in Algorithm 6. Initially the weights of the neurons were randomly set to zeros and ones. Then the best matching unit (BMU) of each input vector was calculated using (5). As can be seen in (7) any feature with the missing value was considered as a matching case with any weight value.

$$BMU^j = argmin_m D(\mathbf{x}^j, \mathbf{w}^m) \tag{5}$$

where

$$D(\mathbf{x}^j, \mathbf{w}^m) = \sum_{i=1}^{i=N} \delta(\mathbf{x}_i^j, \mathbf{w}_i^m) \tag{6}$$

$$\delta(\mathbf{x}_i^j, \mathbf{w}_i^m) = \begin{cases} 0 & if(\mathbf{x}_i^j = \mathbf{w}_i^m) \\ 0 & if(\mathbf{x}_i^j = -1) \\ 1 & if(\mathbf{x}_i^j \neq \mathbf{w}_i^m) \end{cases} \tag{7}$$

Once the BMU of each input vector was calculated the weights of each neuron were updated using (8). As seen in (8), the weights get the value of the most frequent categorical value of each feature (ignores the categorical value of missing data). The function $h_{BMU^j m}$ is a Gaussian neighbourhood function centred at the BMU of $\mathbf{x}^j$.

$$\mathbf{w}_i^m(t+1) = \begin{cases} \mathbf{a}_c & if(F(\mathbf{a}_c, \mathbf{w}_i^m(t)) > F(\mathbf{a}_{r \notin \{c,3\}}, \mathbf{w}_i^m(t))) \\ \mathbf{a}_c & if(F(\mathbf{a}_c, \mathbf{w}_i^m(t)) = F(\mathbf{a}_{r \notin \{c,3\}}, \mathbf{w}_i^m(t)) \\ & \wedge\, random(0,1) > 0.5) \\ \mathbf{w}_i^m(t) & otherwise \end{cases} \tag{8}$$

where

$$F(\mathbf{a}_r, \mathbf{w}_i^m) = \frac{\sum\limits_{j=1}^{j=P} (h_{BMU^j m} | \mathbf{x}_i^j = \mathbf{a}_r \vee \mathbf{x}_i^j = -1)}{\sum\limits_{j=1}^{j=P} h_{BMU^j m}},\ r = 1, 2 \tag{9}$$

$$c = arg_r max F(\mathbf{a}_r, \mathbf{w}_i^m),\ r = 1, 2 \tag{10}$$

Once the weights were updated, the neighbourhood ($radius$) was decreased. Then the process was repeated for a predefined number of epochs T.

5.2.4 Models Investigated

Different map sizes were evaluated on the input dataset using a 10-fold cross validation. Each fold had the same proportion of cases and controls based on the initial distribution of cases and controls of the dataset. Specifically each fold contained approximately 162 cases and 199 controls and the total number of subjects used in training and testing are shown in Table 17. Due to the initial randomization of the weights of SOM, the 10-fold cross validation procedure was repeated 10 times. To select the appropriate map size for SOM, the test was repeated on a 2x2, a 4x4, a 6x6 and an 8x8 map size for 1000 epochs. The initial neighbourhood radius was set as half of the map's edge size.

5.2.5 Evaluation Metrics

5.2.5.1 Traditional Measures

To evaluate the classification power of SOM and test how the map size of the network affects the results, each cluster of the trained network was labeled as ”case” if the majority of its subjects were cases and ”control” otherwise.

Table 19: Confusion Matrix

		Actual	
		Control	Case
Predicted	Control	TP	FP
	Case	FN	TN

Then the confusion matrix was calculated as shown in Table 19 and the following evaluation metrics were calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{14}$$

5.2.5.2 Pearson's Chi Square Test (χ^2)

Because the evaluation metrics mentioned above do not take into consideration the map size, χ^2 was also used for evaluating the clustering of the network. The hypothesis here is that the generated clusters will be associated with the case/control status. The χ^2 provides a well established methodology to test that. Specifically the null hypothesis tested, is that there is no association among the clusters generated and the case/control status of the subjects in the clusters.

For testing this, two contingency tables were created after the training of each SOM run. The first had the number of cases and controls of each neuron (cluster) using the training data and the second used the testing data. Then the p-value of χ^2 was calculated on each contingency table. It was decided a-priori that the null hypothesis would be rejected if the p-value was smaller than 0.01 $(-log(p\text{-}value) > 2)$. The advantage of this test is the fact that the p-value calculated takes into consideration not only the distribution of the cases and controls in each cluster, but also the number of active clusters. An active cluster in this work is defined as a cluster with at least one subject assigned in it.

Table 20: Clustering Results Using the Top 37 SNPs in the HLA Region

		SOM map size			
		2x2	4x4	6x6	8x8
χ^2 $(-log(p\text{-}value))$	Train	21.0±2.1	**42.9**±2.9	39.9±3.0	35.1±2.5
	Test	2.7±1.4	**3.1**±1.4	2.0±1.0	1.7±0.4
Accuracy (%)	Train	58±0.3	63±0.7	63±0.7	**64**±0.8
	Test	58±3	62±2.7	63±2.3	**63**±2.2
Sensitivity (%)	Train	64±2.0	66±3.5	70±1.9	**70**±1.7
	Test	64±4	66±4.2	**69**±2.9	69±3
Specificity (%)	Train	51±2.7	**58**±4	56±2.6	57±2.5
	Test	51±5	**58**±6.4	55±5.4	55±5.4
Balanced Accuracy (%)	Train	58±0.4	62±0.8	63±0.8	**64**±0.8
	Test	57±2.6	62±2.8	62±2.5	**62**±2.4

5.2.6 Pattern Identification

After training, the best performing model was selected. The parameters of this model (e.g. map size, neighbourhood radius) were used for constructing an SOM, where the whole dataset was used for training. Once the model was trained, the generated clusters were tested for association with the case-control status of the subjects using χ^2. The χ^2 was calculated using a contingency table with the number of cases and controls of each cluster. If the statistical significance of the association was above the predefined threshold ($-log(p\text{-}value) \geq 2$), the patterns of the clusters would be further investigated to identify SNP associations with the disease status of the subjects clustered. Since each weight of a cluster represents the allelic value of the majority of the subjects of that cluster, the weights of the trained SOM were used to identify any interesting patterns among SNPs. To have more representative results (since the weight values were dependent on the distribution of the two classes in a cluster), only clusters with a minimum separation of 70% - 30% among the two classes were selected for pattern identification.

5.3 Results

5.3.1 Models Investigated

A total of 37 SNPs were selected by the two SNP interaction algorithm [101] and these SNPs were used as input for all tests performed. The results of the 10-fold cross validation for both the train and test sets are presented in Table 20. From the χ^2 test, it is clear that the best map size for the selected dataset is the 4x4 map size. Specifically there is a major increase in the $-log(p\text{-}value)$ from the 2x2 size to the 4x4 size and then it decreases as the map size increases. The $-log(p\text{-}value)$ was 42.9 ± 2.9 for the training set and 3.1 ± 1.4 for the testing set, which are above the predefined threshold. For the testing set, the measure was close to the predefined threshold, but this was mainly because of the small number of subjects used in the testing phase. To address this, models were retrained using half of the subjects for training and the other half for testing on 5 different such sets. SOM was run 10 times on each set using the 4x4 map size and statistically significant clusters were obtained for both training and testing results with a $-log(p\text{-}value)$ close to 21, which is far greater than the predefined threshold.

The percentage of balanced accuracy was 62 ± 2.8 for the test set of the 4x4 map size. Similar to this value but with slightly smaller standard deviations were obtained for the 6x6 and 8x8 map sizes. The percentage of sensitivity and specificity were 66 ± 3.5 and 58 ± 6.4 for the test set for the 4x4 map size. Similar values were also obtained for the 6x6 and 8x8 map sizes.

5.3.2 Pattern Identification

After training the SOM with a 4x4 size map using all of the subjects and the top 37 SNPs, the $-log(p\text{-}value)$ was 45, indicating that the weights of its clusters were associated with the disease. In this model, 5 major clusters were identified with more than 65% separation among the normalized distribution of cases and controls. Two of them are ”case” clusters and the others are ”control” clusters. In total these 5 clusters account for 1312 subjects, which is approximately 36% of the total number of subjects. These 5 clusters are highlighted with a grey background in Figure 8. In each pie chart there are two numbers where the number in the blue area represents the fraction out of the total controls that got clustered in that specific cluster and similarly the same applies for the number in the red area and cases.

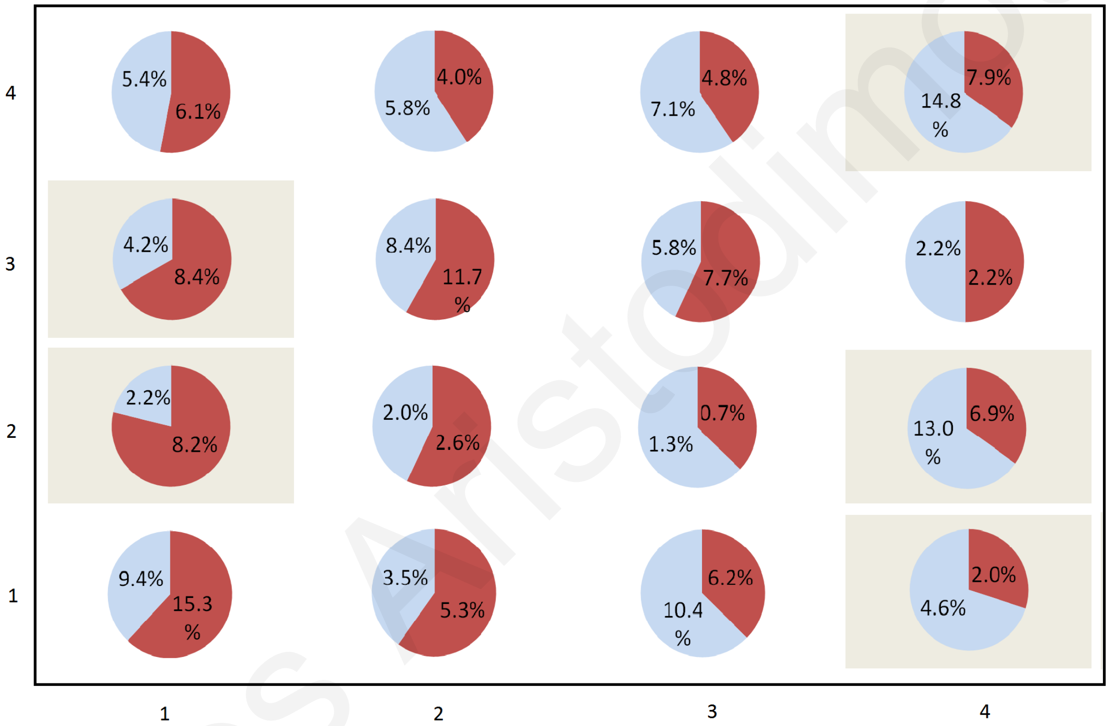


Figure 8: The normalized distribution of cases and controls in each cluster after training SOM on a 4x4 map using all of the subjects for training with the top 37 SNPs. Light blue represents the controls and red represents the cases. The values in the blue areas in the pie charts represent the proportion out of the total controls that are in each cluster and similarly the numbers in the red areas the proportions out of the total cases. Highlighted with a grey background are the top 5 clusters based on the difference of their distribution of cases and controls.

Out of these 5 clusters, 4 of them had a proportion of cases and controls above the a-priori defined threshold (70% - 30%). In Figure 9 each row illustrates the weights of these 4 clusters. The numbers in the first column represent the (x,y) coordinates of each cluster as illustrated in

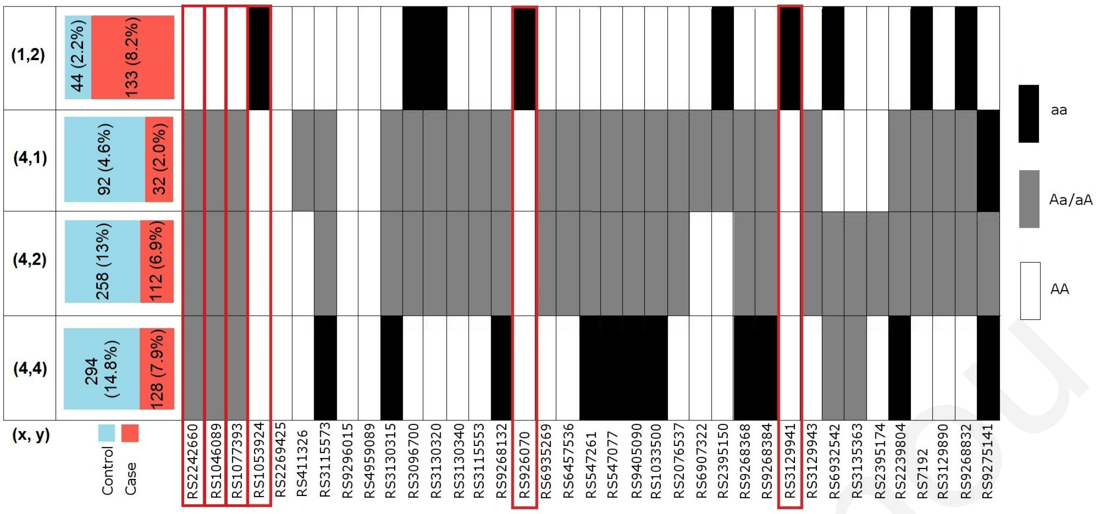


Figure 9: Pattern identification showing the top 4 clusters of a 4x4 map using the top 37 SNPs as input. The numbers in the first column represent the (x,y) coordinates of each cluster as illustrated in Figure 8. The second column has bar charts with details such as the number of cases and controls in each cluster and the fraction out of the total number of cases and controls that each number corresponds in parentheses. The rest of the columns have the weight values of the clusters, that represent the most frequent allele values of the subjects of that cluster for the specific SNP. ”AA” represents homozygous major allele value, ”aa” homozygous minor allele value and ”aA/Aa” heterozygous allele value for a SNP. The highlighted columns (rs2242660, rs1077393, rs1046089, rs1053924, rs926070, rs3129941) show important differences in patterns among the case and controls clusters.

Figure 8. The second column has bar charts showing the distribution and the number of cases and controls in each cluster. The numbers in parentheses in the bar charts, represent the fraction out of the total number of cases and controls for each cluster. The rest of the columns represent the weight values of the two alleles for each SNP. As can be seen, the ”control” clusters have similarities among them that are not present in the ”case” cluster. Three SNPs that illustrate this are rs1053924, rs926070 and rs3129941, which are highlighted with red rectangles. In these three SNPs, the majority of the subjects in the case cluster had homozygous minor allele values whereas in the controls clusters, the majority had homozygous major allele values. Similarly for the first three SNPs (rs2242660, rs1077393, rs1046089) in Figure 9, the case cluster had homozygous major allele values, whereas the control clusters had heterozygous allele values. The two SNP

interaction algorithm indicated for these three SNPs that rs2242660 interacted with rs1077393 and that rs1077393 interacted with rs1046089, hence SOM identified a three SNP interaction among these SNPs.

5.4 Discussion

In this chapter, SOM, a clustering algorithm, was tested to investigate whether it could find clusters whose SNPs were strongly associated with MS. From the results obtained, an important finding is that the clusters identified are statistically significant. The top clusters had an important separation among cases and controls and they accounted for a good proportion of the total subjects of the dataset. Moreover, some interesting patterns among the top case and controls clusters were identified. Further investigation of these patterns needs to be performed for identifying the actual causative effects driving them.

Many studies indicated that there is an association among the Human Leukocyte Antigen (HLA) region and MS using single locus association testing [97, 98, 99, 100]. Antoniades used a two SNP interaction algorithm in [101] and was able to identify statistically significant two SNP interactions among 37 SNPs, which were replicated using an independent dataset. Those 37 SNPs were used and SOM was able to identify higher order SNP interactions, with 6 SNPs revealing interesting patterns among the ”case” and ”control” clusters. Brassat *et al* identified single and three locus association models among SNPs in the HLA region and MS using MDR [103], whereas in [104], Motsinger *et al* identified two, three and four locus associations among SNPs in the HLA region and MS using MDR as well. The SNPs identified by these two studies were not selected by the feature selection algorithm used in this work, hence the associations identified cannot be compared with the associations identified in those two studies. These findings indicate that SOMs can be used for clustering GWA data for finding associations among SNPs and a disease. The advantage of the SOM is that it is searching for n-SNP associations when creating the clusters. This is accomplished by clustering subjects together that have as many similar SNP patterns as possible without performing an exhaustive search as MDRs [76].

From the metrics used for evaluating the clustering and selecting the map size of the network, the χ^2 metric was more indicative than the traditional evaluation metrics. It has the advantage of considering the initial distribution of the classes, which is important when analysing imbalanced data. This is something that the standard accuracy measure does not cope with, making it inappropriate for such cases. Balanced accuracy can be used with imbalanced data and when used with specificity and sensitivity, the accuracy on each class and their average accuracy can be observed. But these measures do not take into consideration the number of clusters used by the network as the calculation of the p-value of χ^2 does. Moreover the p-value calculated accounts for the number of subjects used as well, hence χ^2 can be used as a single measure for evaluating and selecting the map size of the network.

5.5 Conclusion

The ability of SOM for finding associations among SNPs and a disease has been investigated with promising results. From the results obtained, it can be seen that the unsupervised clustering of SOM was statistically significant, revealing an association among the SNP patterns in the clusters generated and the disease status of the subjects. Moreover the χ^2 statistic due to its ability of taking into consideration the distribution of the classes, the number of subjects in the dataset and the number of active clusters of the network, has been proposed for selecting the map size of the network instead of the traditional evaluation measures. Finally we conclude that SOMs ability of clustering subjects with similar SNP patterns together, is an important feature that may prove of significant value in future multi-loci association testing.

Chapter 6

A Framework for Efficient n-Way Interaction Testing in Case/Control Studies

6.1 Introduction

The introduction of affordable high throughput genotyping technologies allows the assay of millions of genetic polymorphisms per subject across the whole genome. This has led to genome wide association studies (GWAS) that try to identify genetic variations associated with diseases. Traditional genetic analyses focus on single locus associations, but most common diseases are influenced by multiple gene interactions and interactions with the environment [5].

Gene interaction is known in biology as epistasis, which is the result of physical interactions among biomolecules within gene regulatory networks and biochemical pathways in an individual, such that the effect of a gene on a phenotype is dependent on one or more other genes [105]. There is also the term of statistical epistasis, which is defined as the deviation from additivity in a statistical model, summarizing the relationship between multi-loci genotypes and phenotypic variations in a population [105].

Various methods have been proposed for interaction testing such as statistical [106], information theory [107, 108] and machine learning [76, 109] based methods. The statistical and information theory methods, usually provide a measure for identifying an interaction, but one needs to perform an exhaustive search on the available variables. To test for all possible n-Way interactions, one would need $O(k^n)$ tests, where k is the total number of variables. This increases the computational complexity of the analysis. Additionally, in statistical tests that compute a p-value, if Bonferroni correction is used to address the problem of multiple testing, many true positive interactions will be missed, because of the conservative nature of this adjustment and the large number of tests performed [109].

Machine learning algorithms have also been proposed for identifying interacting factors. The Multifactor Dimensionality Reduction (MDR) algorithm [77], has been used in many studies [110, 111, 112, 113]. This method is model free and reduces the dimensionality of tested factors by labeling each pattern as ”high risk” or ”low risk” if the ratio of cases:controls is above a threshold. Then using these labels and Cross Validation (CV), the model is evaluated. Random Forest (RF) [23] based methods have also been proposed for interaction testing [89, 88], but have not been as widely used as the MDR based methods.

Machine learning algorithms for pattern recognition can also be used for identifying interesting patterns associated with an outcome. The features involved in these patterns can then be used to test if interactions also exist. This could involve association rule mining alogorithms [114, 24, 115, 116] that try to identify frequent patterns in the data, and clustering algorithms [117, 90, 118] that try to cluster instances in different groups based on their features similarity. In clustering algorithms, one can see the most frequent values of each cluster and the class they represent and see if there are any interesting patterns that are different between the classes and focus on those [4].

Applying machine learning algorithms directly on high dimensional data can be computationally demanding and can result in models that are hard to interpret [119]. Additionally in the case that the number of features is much larger than the number of instances, a common case in biomedical data, it is hard to identify the true signals from noise and there is also the risk of overfitting [120]. To reduce the search space before performing interaction testing and to also alleviate the issues aforementioned for machine learning methods, feature selection can be used [121, 119]. Feature selection, tries to identify the smallest subset of the available features that can best explain the response variable. Various feature selection algorithms have been used in genetic data such as Relief-F [122], the Recursive Feature Elimination Support Vector Machine (RFE-SVM) [123], and various penalized regression methods [**?**].

In this chapter we propose a machine learning framework for discovering n-Way interactions associated to disease that is demonstrated on real data from a Multiple Sclerosis study. The goal is to use machine learning to reduce the dimensionality of the data and therefore limit the multiple testing problem and the complexity of the analysis. Then subjects are clustered based on their underlying genetic profile across multiple genetic loci. Finally, using the most frequent genotypes of the clusters, the genotypes are converted to binary to reduce the degrees of freedom of the interaction tests that will be performed and increase the statistical power.

6.2 Proposed Framework

The proposed framework is comprised of four steps. The first step is on quality control to remove any erroneous data. The next two steps use machine learning to perform feature selection and to cluster subjects based on their similarities. The final step uses the processed data of the previous steps to convert each feature into a binary variable before performing an n-Way interaction testing. Essentially, the last step reduces the dimensionality of each feature and hence the degrees of freedom of the statistical test that will be performed. This, increases the statistical power of the analysis. The proposed framework can be applied on categorical data and hence any continuous features that need to be included in the analysis should be discretized. The features can be of any type (e.g. genetic, environmental, phenotypic), which means that the analysis is not limited in the identification of only interacting genetic data, but also the interactions between genetic and environmental/phenotypic data.

6.2.1 Quality Control

The purpose of the first step is to perform data cleaning to remove data of low quality that could negatively affect the analysis. In this step, one can use any of the well established methods for quality control in genetic data. Some of these are the Hardy-Weinberg Equilibrium (HWE) test, the minimum Minor Allele Frequency (MAF) filter and the percentage of missing values filter.

6.2.2 Feature Selection

Once data have been cleaned, supervised feature selection is used. Since the ultimate goal is to identify interactions, it is important to use an algorithm that can select such interacting features. This means that it is better to avoid using feature selection algorithms that perform univariate analysis, since this can exclude interacting terms that have low main effects. Another important characteristic of the feature selection algorithm that will be used, is to use a similar modeling function as the interaction testing method. For example, if the interaction testing method is model-free, then it is best if the feature selection algorithm is also model-free, otherwise important features could be removed.

6.2.3 Clustering

The next step is to cluster subjects based on their genotype. The similarity between data is usually calculated using the Hamming distance when they are categorical. In this framework, it is of interest to group subjects that have similar genotypic/environmental/phenotypic patterns together and use those to help perform a more targetted interaction testing. For this reason, features are handled as nominal categorical variables in the clustering procedure, since we do not want to consider an order in the values of each feature. Once clustering is performed, each resulted cluster will have a different number of cases and controls. If the majority of the subjects in that cluster are Cases then the cluster will be labeled as a Cases cluster otherwise as a Controls cluster. Based on these requirements, it is important to use a clustering algorithm that can handle nominal categorical data such as the NC-SOM [95] and k-modes [124].

Finally, to select the appropriate number of clusters and to see if the resulting clustering is of interest, the clusters should be evaluated. This can be tested using Pearson's χ^2 test on the number of cases against the number of controls in each cluster, as was proposed in [4]. Using this test, one can see if the distribution of cases and controls in the produced clusters deviates from the expected and hence indicates that the patterns in the clusters could be used to separate the two classes.

6.2.4 n-Way Interaction Testing

In the last step, given that the clustering had statistically significant results, the features will be encoded to binary variables. This is done by initially identifying clusters that are of interest, based on the ratio of cases and controls, the size of the clusters and how distinctive the patterns of a cluster of a certain label are from the patterns of clusters of other labels.

Once a cluster of interest is identified, the features' values are encoded into binary, based on the most frequent value the subjects of that cluster have for that feature. For example, if we had SNPs and for an SNP, the subjects in the cluster of interest mainly have heterozygous alleles then aB and Ba will be encoded to the value of 1 and the homozygous minor alleles (aa) and homozygous major alleles (BB) will be encoded to 0. Then n-Way interaction testing is applied to all subjects using the new encoding.

This allows performing targeted interaction testing, since the clustering already identified the values that are common in the cluster of a certain class. Hence, we are now investigating if these patterns also have an interacting effect. Additionally, since we have reduced the dimensionality of each feature to binary, the statistical power of the analysis is increased, which allows identifying such interactions with smaller sample sizes as well.

6.3 Methodology

The dataset used in the experiments is presented along with the evaluation methodology of the proposed framework. For each step of the proposed framework, the algorithms used are extensively described.

6.3.1 Dataset

The distribution of the subjects in the dataset used in this study is shown on Table 21. It consists of 389 Multiple Sclerosis (MS) patients and 336 controls from 3 MS centers; The Cyprus Institute of Neurology and Genetics in Cyprus, the University Hospital of Larissa, and the AHEPA Hospital of Aristotle University.

Table 21: Subjects Distribution

	Males	Females	Total
Cases	138	251	389
Controls	102	234	336
Total	240	485	725

All patients are of Greek origin with 240 of them being males and 485 of them being females. The dataset has 147 tagging SNPs across 9 genes encoding for P-selectin (SELP), integrins (ITGA4, ITGB1, and ITGB7), adhesion molecules (ICAM1, VCAM1, and MADCAM1), fibronectin 1 (FN1), and osteopontin (SPP1) [125]. The genotypes of the SNPs were assigned with the value of 1 for homozygous minor alleles (aa), the value of 2 for heterozygous alleles (aB/Ba) and the value of 3 for homozygous major alleles (BB). Missing values were represented by the value of 0.

6.3.2 Evaluation Methodology

This section outlines the methods used in this study in each step of the proposed framework. Additionally the methodology followed for evaluating the effectiveness of the framework is described.

6.3.2.1 Quality Control

For data cleaning, two data characteristics were used. The first one was the percentage of missing data of each SNP, and the second one the MAF. Specifically, any SNPs with more than 5% missing values or with an MAF less than 5% were removed from the dataset.

6.3.2.2 Feature Selection

For features selection, Random Forest (RF) [23] was used for finding the importance of the SNPs regarding their ability of differentiating between cases and controls. RF is a collection of many de-correlated trees and to decide in which class a subject belongs to, each tree votes and the class with the majority of votes is selected. The algorithm creates each tree using bootstrapping hence at each tree some subjects will be used more than once, whereas some others will not be.

The subjects not used are called the out of bag (OOB) samples and are used for calculating the misclassification error.

In this work, the implementation from the R package randomForest was used [48] using 50 trees and the importance of SNPs was measured by the Mean Decrease in Gini.

6.3.2.3 Clustering

For clustering, the NC-SOM [95] was used and specifically the version for nominal categorical data with the methodology described in [4] for selecting the appropriate map size. SOMs have the ability of assigning subjects with similar genetic data to neighboring clusters. Hence clusters that are close together tend to have subjects that are similar to each other compared to subjects in clusters in further locations. The distance between a cluster and a subject is calculated using the Hamming distance, which is more appropriate for nominal categorical data. Additionally the weights of each cluster, which represent the center of the cluster, are the modes of each cluster and hence one can see the most frequent values of each SNP for that cluster. One difference to the encoding applied to the data in [4], is that in this paper, SNPs are encoded to the genotype level with the values $\{0, 1, 2, 3\}$ for $\{missing, aa, aB/Ba, BB\}$ respectively.

Different map sizes are evaluated on the input dataset using a stratified 2-fold Cross Validation (CV). Due to the initial randomization of the weights of SOM, the 10-fold cross validation procedure is repeated 10 times. To select the appropriate map size for SOM, the test was repeated on different map sizes for 1000 epochs. The initial neighbourhood radius is set to half the map's edge size. The χ^2 test is used for evaluating the clustering on both training and testing data, by applying it on the contingency table that has the number of cases and controls of each cluster. The map size with the lowest p-value is selected and if it is statistically significant $(p - value <= 0.05)$ the clustering is repeated with all data and the analysis proceeds to the next step of the framework.

6.3.2.4 n-SNP Interaction Testing

For interaction testing, two methods were used to test if the proposed framework behaves similarly in different methods. The first method is the classic logistic regression and the second a χ^2 based interaction testing method [126]. Both methods were testing for 2-SNP and 3-SNP interactions.

For the logistic regression, the model used for 2-SNP interaction testing is shown in (15)

$$ln(\frac{p}{1-p}) = \beta + \beta_A x_A + \beta_B x_B + \beta_{AB} x_A x_B \tag{15}$$

in which x_A and x_B are the two SNPs of interest at loci A and B respectively, β_A and β_B are regression coefficients that represent the main effects of exposures at A and B, and the coefficient β_{AB} represents an interaction term [75]. Similarly the model can be extended for 3-SNP interactions as shown in (16).

$$\begin{aligned} ln(\frac{p}{1-p}) &= \beta + \beta_A x_A + \beta_B x_B + \beta_C x_C + \beta_{AB} x_A x_B \\ &+ \beta_{AC} x_A x_C + \beta_{BC} x_B x_C + \beta_{ABC} x_A x_B x_C \end{aligned} \tag{16}$$

The model for 2-SNP interaction testing using the χ^2 method is shown in (17)

$$\chi^2_{A*B} = \chi^2_{A+B} - \chi^2_A - \chi^2_B \tag{17}$$

where χ^2_{A+B} is the omnibus effect of SNPs A and B, χ^2_A is the main effect of SNP A and χ^2_B is the main effect of SNP B. To calculate the degrees of freedom (18) is used

$$DF_{A*B} = DF_{A+B} - DF_A - DF_B \tag{18}$$

where DF_{A+B} is the degrees of freedom of the χ^2 test performed on the omnibus effect of SNPs A and B and similarly DF_A, DF_B for the main effect of SNPs A and B respectively. To test for 3-SNP interactions the model is extended as shown in (19)

$$\begin{aligned}\chi^2_{A*B*C} = \chi^2_{A+B+C} - \chi^2_{A*B} - \chi^2_{B*C} - \chi^2_{A*C} \\ - \chi^2_A - \chi^2_B - \chi^2_C\end{aligned} \tag{19}$$

and its degrees of freedom can be calculated using (20)

$$\begin{aligned}DF_{A*B*C} = DF_{A+B+C} - DF_{A*B} - DF_{B*C} - DF_{A*C} \\ - DF_A - DF_B - DF_C\end{aligned} \tag{20}$$

Interaction testing was performed on both the binary version of the genotypes and on the initial encoding, to test if the proposed encoding helps with the task at hand. An exhaustive 2-SNP and 3-SNP interaction testing was performed on the selected SNPs. In each tested SNPs pair/triplet, the subjects that had a missing value in the SNPs of that pair/triplet, were removed from the analysis. An interaction was considered statistically significant if it had a p-value less than 0.05. To address the multiple testing problem, permutation analysis was performed to adjust the p-values using 10,000 permutations.

6.4 Results

From quality control, 80 SNPs failed to pass the missing data threshold and from the remaining ones two more failed to pass the MAF threshold. Hence the Random Forest had 65 SNPs to analyze. The SNPs importance plot is shown in Figure 10.

Based on the mean decrease in Gini, the top 7 SNPs of Figure 10 were selected. All 7 SNPs were also found having an association with the case/control status of the subjects in [125]. This

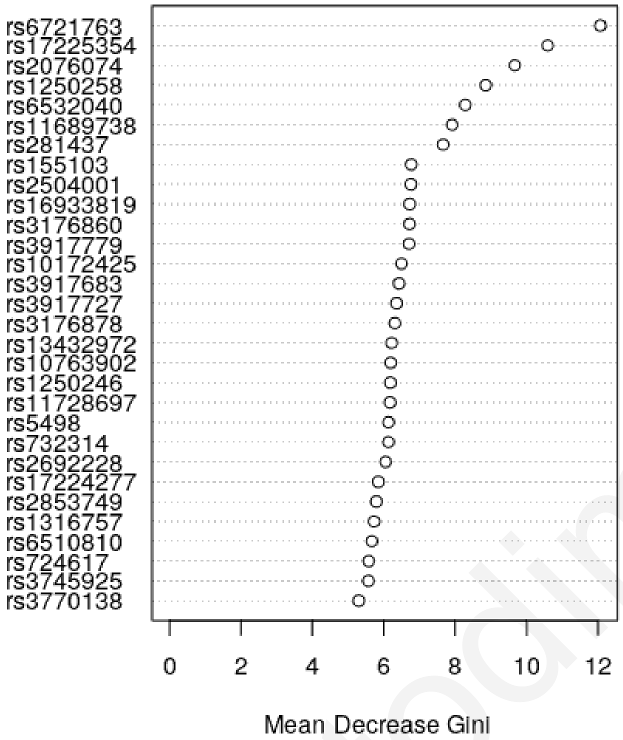
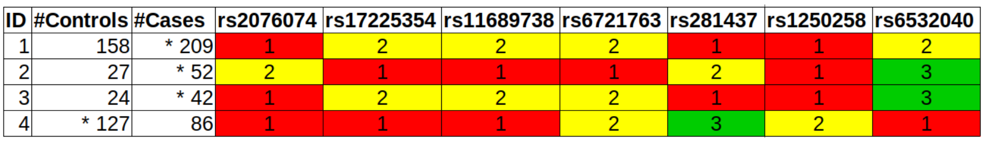


Figure 10: The importance of the top 30 SNPs from the Random Forest analysis using the mean decrease in Gini.

means that in the selected SNPs there are no SNPs with a low main effect and that any possible interactions in this case are between SNPs that have high main effects. Random Forests can include SNPs that have low main effects if they are interacting with SNPs with high main effects, but if the interacting SNPs are all with low main effects it is possible that they will be missed [76], hence such interactions might have been missed.

After testing different map sizes for the SOM using cross validation, the final topology was selected to be a 2x2 map. The clustering with all subjects using the selected SNPs from Random Forest had a p-value of 0.00002, indicating that the distribution of cases and controls in the clusters was deviating from the expected. The final clusters are shown in Figure 11.

Each row represents a cluster and has its identification number (ID) in the first column and the number of cases and controls in the next two columns. An asterisk indicates the majority class



ID	#Controls	#Cases	rs2076074	rs17225354	rs11689738	rs6721763	rs281437	rs1250258	rs6532040
1	158	* 209	1	2	2	2	1	1	2
2	27	* 52	2	1	1	1	2	1	3
3	24	* 42	1	2	2	2	1	1	3
4	* 127	86	1	1	1	2	3	2	1

Figure 11: Each row represents a cluster of the SOM. The first two columns indicate the number of cases and controls each cluster has and an asterisk indicates if the majority of subjects is cases or controls. The rest of the columns show the most frequent genotype of each SNP for the subjects that belong to it.

of a cluster. Then for each SNP the most frequent genotype of the subjects that belong to that cluster is shown. Red indicates homozygous minor alleles, yellow heterozygous alleles and green homozygous major alleles. As indicated on Figure 11, the first three clusters (IDs 1-3), include a majority of cases in them and only the last cluster (ID 4) has a majority of controls. The encoding of the SNPs to binary, will be based on the controls cluster to evaluate if there are SNP interactions with those genotypes that can help separate cases from controls. For example, SNP rs281437 will have the value 1 for its homozygous major alleles and 0 for the rest.

The interaction testing results are shown in Table 22. Both logistic regression and the χ_2 method identified more 2-SNP and 3-SNP interactions using the proposed binary encoding compared to using the initial encoding. In both methods methods there a single 2-SNP interaction that was only identified using the initial encoding, whereas using the binary encoding the logistic regression identified four more SNP interactions and the χ^2 method five more that were not identified using the initial encoding.

Table 22: Statistically Significant n-SNP Interactions per Method and SNP Encoding

	Initial Encoding	**Binary Encoding**	**Initial Encoding**	**Binary Encoding**
2-SNP interactions	**Logistic**		χ^2	
rs17225354 (ITGA4) & rs6721763 (ITGA4)	✓		✓	
rs17225354 (ITGA4) & rs281437 (ICAM1)	✓	✓		✓
rs17225354 (ITGA4) & rs1250258 (FN1)		✓		✓
rs2076074 (SELP) & rs6721763 (ITGA4)		✓		✓
rs11689738 (ITGA4) & rs281437 (ICAM1)	✓	✓	✓	✓
rs11689738 (ITGA4) & rs6721763 (ITGA4)		✓		✓
3-SNP interactions	**Logistic**		χ^2	
rs17225354 (ITGA4) & rs281437 (ICAM1) & rs1250258 (FN1)	✓	✓		
rs2076074 (SELP) & rs6721763 (ITGA4) & rs281437 (ICAM1)		✓		✓

There are two reasons for being able to identify more n-SNP interactions using the proposed encoding. Initially, with binary encoding, the degrees of freedom of the statistical tests are reduced and hence the power of the analysis is increased. For example, in the χ^2 method, using the initial encoding in the 2-SNP interaction test we ended up with four degrees of freedom, whereas using the binary encoding with only one. Similarly, for the 3-SNP interaction test, we dropped from 8 degrees of freedom to one. The second reason is due to the fact that the encoding enforces a more targeted analysis on the possible interacting SNP values that can be seen from the clustering step, which shows the SNP genotypes that are more frequent in the cluster of interest versus the rest, which in this case is the controls cluster versus the cases clusters.

6.5 Conclusion

A framework for the efficient identification of n-Way interactions has been proposed. The framework removes any erroneous data in its first step and then uses feature selection to reduce the dimensionality of the search space and select the most informative features. This, reduces the statistical tests that will be performed and hence the multiple testing problem. Additionally, it reduces the computational complexity of the analysis to be performed and the probability of overfitting. Therefore, the first two steps of the framework focus on providing high quality and informative features.

In the clustering step, subjects are grouped based on their similarity in their genotypes and each cluster is labeled based on the majority class of the subjects. This enables the visualization of the most frequent values of the subjects in each cluster and the phenotype they represent. Then a cluster of interest is selected and its values are used to convert the genotypes to binary variables. This enables a targeted interaction testing analysis and additionally increases the power of the statistical analysis due to the reduction of the degrees of freedom of the tests. As has been shown in the results, using the binary encoding more 2-SNP and 3-SNP interactions were identified, which further supports the hypothesis that the proposed framework helps identifying more interactions.

Additionally, due to the cluster based binary encoding, the framework can also promote personalised medicine. For example, one could identify different n-Way interactions for each cluster of interest and hence adjust the treatment of a patient based on his/her genotypic profile.

Chapter 7

Concluding Remarks

7.1 Conclusion

This work began with the project Linked2Safety, a semantically interconnected approach for sharing aggregated data in the form of data cubes, which preserves individuals' privacy, while enabling the multi-source, multi-type analysis of health data through a single web based secure access platform. From the experience acquired in the project, two problems were identified that needed to be automated. The first one is the need for automating the selection of the most informative variables to be published in the platform so as to increase the probability of significant findings in the analyses that would be performed. The second one was the need for converting any continuous variables into categorical to be able to use the data cube approach with cell suppression for anonymizing the data.

To address the problem of selecting the most informative features and anonymizing data, kPB-MS was proposed. The algorithm combines feature selection with k-anonymity, so as to select the most informative features that conform to the k-anonymity criterion and to reduce the number of instances removed from the dataset. For the feature selection algorithm, a novel measure was proposed which is model free and computationally efficient so as to be able to capture any type of variable associations with the target variable and to be able to analyze efficiently high dimensional data. The multidimensional suppression procedure of kPB-MS takes into consideration the instances removed and the effect of record suppression on classifiers. The first is obtained by allowing the user to define the accepted loss of instances, whereas the second is obtained by testing the significance of the suppression using Fisher's exact test on each pattern. As shown in the results, the algorithm did not negatively affect the classifiers in 80% of the test cases, indicating that kPB-MS can be used in privacy preserving data publishing.

The second problem was addressed with the proposal of DBAD, a novel supervised discretization algorithm, based on how the density of the values of a feature changes for each class. DBAD had similar or better results with the state of the art discretization algorithms that were used in the evaluation procedure in the majority of the tests, showing that density based discretization is a promising approach and needs to be further investigated. DBAD is computationally efficient and embarrassingly parallel. The parallel implementation of the algorithm showed that it can obtain a linear speedup, which makes it appropriate for use on big data.

To identify n-SNP patterns associated with the disease status of individuals, a categorical version of SOM was used. For selecting the map size of the network, the χ^2 statistic was proposed, due to its ability of taking into consideration the distribution of the classes, the number of individuals in the dataset and the number of active clusters of the network, to avoid overfitting. The produced clustering of SOM was statistically significant, revealing an association among the SNP patterns in the clusters generated and the disease status of the subjects, indicating that clustering subjects with similar SNP patterns together, is an important feature that may prove of significant value in future multi-loci association testing.

Finally, since the identification of interactions is important in common complex diseases, a framework for the efficient identification of n-Way interactions was proposed. The framework removes any erroneous data in its first step and then uses feature selection to reduce the dimensionality of the search space and select the most informative features. Therefore, the first two steps of the framework focus on providing high quality and informative features. The next step uses clustering to group subjects based on their similarity in their genotypes and each cluster is labeled based on the majority class of the subjects. Then a cluster of interest is selected and its values are used to convert the genotypes to binary variables. This enables a targeted interaction testing analysis and additionally increases the power of the statistical analysis due to the reduction of the degrees of freedom of the tests. Results indicate that using the binary encoding we were able to identify a larger number of statistically significant 2-SNP and 3-SNP interactions compared to using the initial encoding of the genotypes, which further supports the hypothesis that the proposed framework helps identifying more interactions. Additionally, due to the cluster based binary encoding, the framework can also promote personalised medicine, since one can identify different n-Way interactions for each cluster of interest and adjust the treatment of patients based on their genotypic profiles.

Overall the four proposed methodologies have shown promising results and can be used in different aspects of the analysis of categorical data in the medical domain. From converting variables into categorical, to performing privacy preserving data publishing so as the data from multiple sources can be combined, and finally to the analysis of such data for the identification of n-Way associations or interactions that can be ultimately used in drug discovery and for adjusting patients lifestyle for reducing the risk of developing a disease.

If anyone is interested in validating these results, the source code of the algorithms used and the experiments performed can be shared upon request along with the publicly available datasets.

For datasets such as the ones on multiple sclerosis from ANZGENE and the Cyprus/Greece collaboration, one should request access from the consortium that owns the datasets to be able to perform any analyses on them.

7.2 Future Work

7.2.1 DBAD

The main issue with the current DBAD implementation is its simple strategy for the initial binning of the data of each class for a feature. Currently Sturge's rule is used to select the number of equal-width bins that will be created. Hence, one optimization that can be tested is the use of a technique that can better capture the different densities and create non-equal-width bins, like TUBE [67] and the diagonally-cut histogram. A recursive DBAD execution on each bin could also be tried to decide if a bin should be further splitted and better identify regions in which densities change.

Furthermore, a metric based on class separability could be used to decide which bins to keep after the created bins of each class are merged. This could reduce the final per feature intervals and help classifiers produce simpler models.

7.2.2 n-Way Interaction Testing Framework

Initially, we plan to collaborate with experts in the domain to see if the biological function of the identified interacting genes can be related to MS. This will be based on literature review on the specific genes (SELP, FN1, ITGA4 and ICAM1) to see what information is available on the way these genes function separately and if they have been identified associated with the risk of having MS in other studies. Additionally, bioinformatics tools for pathway analysis will be used, such as pathvisio [127] and path [128]. Biological pathways can be used to identify the steps in biological processes and hence can help with the biological interpretation and understanding of the interactions between these genes. If biological validation of the findings is achieved, then there will be further support that the proposed framework is capable of identifying true positive n-Way interactions even in small datasets.

The proposed framework, can be applied on other Multiple Sclerosis datasets and see if we can replicate the results of the initial study if the same SNPs are available. That would further support that the patterns identified are significant and that they can be used as risk factors for the disease or even for producing new drugs based on their biological functionality.

It can also be used on datasets with genetic and other markers, such as phenotypes and imaging data and see if there are any combinations of such variables that can help separate cases from controls. In such a case, the following pipeline will be followed:

1. Discretize any non categorical variables.
2. Apply the framework on the discretized dataset.

7.2.3 Deep Learning

Deep learning has had very good results especially in image processing [129, 130] and it is worth testing how it behaves when combined with some of the methods described in this thesis. The first thing that would be interesting to see, is how deep learning and specifically Convolutional Neural Networks (CNN) would perform when image data get discretized using the DBAD algorithm and whether that would help achieve a convergence faster and if its accuracy is positively affected as was the case with other algorithms [50] when they used discretized data, compared to using the original continuous data.

Since neural networks, in general, are hard to interpret [131], they cannot be easily used for n-Way interaction testing and the identification of which features are interacting in a case/control study. Instead, a deep neural network could be used after this analysis is performed, using as input the identified interacting features (generated via the proposed n-Way interaction testing framework) to classify the instances into cases and controls. Moreover, such a model could be used for predicting the risk of subjects having a specific disease based on their genetic/environmental/phenotypic profile.

Bibliography

[1] A. Antoniades, C. Georgousopoulos, N. Forgo, A. Aristodimou, F. Tozzi, P. Hasapis, K. Perakis, T. Bouras, D. Alexandrou, E. Kamateri, E. Panopoulou, K. Tarabanis, and C. Pattichis, "Linked2Safety: a secure linked data medical information space for semantically-interconnecting EHRs advancing patients safety in medical research," in *IEEE 12th International Conference on Bioinformatics and Bioengineering (BIBE)*, Larnaka, Cyprus, Nov. 2012, pp. 517–522.

[2] A. Aristodimou, A. Antoniades, C. Georgousopoulos, N. Forgó, A. Gledson, P. Hasapis, C. Vandeleur, K. Perakis, R. Sahay, M. Mehdi, C. A. Demetriou, M.-P. F. Strippoli, V. Giotaki, M. Ioannidi, D. Tian, F. Tozzi, J. Keane, and C. Pattichis, "Advancing clinical research by semantically interconnecting aggregated medical data information in a secure context," *Health and Technology*, vol. 7, no. 2, pp. 223–240, Nov 2017.

[3] A. Aristodimou, A. Antoniades, and C. S. Pattichis, "Privacy preserving data publishing of categorical data through k-anonymity and feature selection," *Healthcare Technology Letters*, vol. 3, no. 1, pp. 16–21, 2016.

[4] A. Aristodimou, A. Antoniades, and C. Pattichis, "Clustering subjects in genetic studies with self organizing maps," in *IEEE 12th International Conference on Bioinformatics and Bioengineering*, Larnaka, Cyprus, Nov. 2012.

[5] J. H. Moore and M. D. Ritchie, "The challenges of whole-genome approaches to common diseases," *JAMA: the journal of the American Medical Association*, vol. 291, no. 13, pp. 1642–1643, 2004.

[6] J. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Human heredity*, vol. 56, no. 1-3, pp. 73–82, 2003.

[7] L. Kun, R. Beuscart, G. Coatrieux, and C. Quantin, "Improving outcomes with interoperable EHRs and secure global health information infrastructure," in *29th Annu. Int. Conf. of the IEEE Eng. in Med. Biol. Soc., 2007. EMBS 2007*, Aug. 2007, pp. 6158 –6159.

[8] "Inovative medicine initiatives," Aug 2012. [Online]. Available: http://www.altaweb.it/documents/imi-call-topics-2009_en.pdf

[9] L. He, X. Li, and P. Huang, "Sharing of ehr clinical test results," *Zidonghuayu Yibiao/ Automation & Instrumentation*, vol. 25, no. 5, pp. 18–21, 2010.

[10] B. A. Stewart, S. Fernandes, E. Rodriguez-Huertas, and M. Landzberg, “A preliminary look at duplicate testing associated with lack of electronic health record interoperability for transferred patients,” *J. of the Amer. Medical Informatics Assoc.*, vol. 17, no. 3, pp. 341–344, May 2010.

[11] O. Kilic and A. Dogac, “Achieving clinical statement interoperability using r-MIM and archetype-based semantic transformations,” *IEEE Trans. on Inform. Technology in Biomedicine*, vol. 13, no. 4, pp. 467 –477, July 2009.

[12] A. Stell, R. Sinnott, and J. Jiang, “A federated data collection application for the prediction of adverse hypotensive events,” in *9th Int. Conf. on Inform. Technology and Applications in Biomedicine, 2009. ITAB 2009*, Nov. 2009, pp. 1 –4.

[13] Y. Xiao, T. Pham, X. Jia, X. Zhou, and H. Yan, “Correlation-based cluster-space transform for major adverse cardiac event prediction,” in *2010 IEEE Int. Conf. on Sys. Man and Cybern. (SMC)*, Oct. 2010, pp. 2003–2007.

[14] N. Ramakrishnan, D. Hanauer, and B. Keller, “Mining electronic health records,” *Comput.*, vol. 43, no. 10, p. 77–81, 2010.

[15] E. Directive, “95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data,” *Official J. of the European Communities*, vol. 281, pp. 31–50, 1995.

[16] R. Faden, T. Beauchamp, and N. King, *A history and theory of informed consent*. Oxford University Press, USA, 1986.

[17] “Privireal: Data protection - greece,” Aug 2012. [Online]. Available: http://www.privireal.org/content/dp/greece.php

[18] “Office of the commissioner for personal data protection - home page,” Aug 2012. [Online]. Available: http://www.dataprotection.gov.cy/dataprotection/dataprotection.nsf/d1813d5911e138bdc2256cbd00313d1c/f8e24ef90a27f34f c2256eb4002854e7

[19] “Federal act on data protection,” Aug 2012. [Online]. Available: http://www.vud.ch/generaldocs/vud_revdsg/235.1_FADP_en.pdf

[20] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, June 2011.

[21] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, “Galaxy: a web-based genome analysis tool for experimentalists,” *Current Protocols in Molecular Biology*, vol. 19, no. 19.10, pp. 11–19, 2010.

[22] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team, “Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences,” *Genome Biology*, vol. 11, no. 8, p. R86, Aug. 2010.

[23] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[24] J. R. Quinlan, *C4. 5: programs for machine learning*. Morgan kaufmann, 1993, vol. 1.

[25] S. Bateman, “Riskmaps: A route to drug safety,” *Good Clinical Practice J.*, vol. 12, no. 9, p. 15, 2005.

[26] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing: A survey of recent developments,” *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, June 2010.

[27] P. Samaratis and L. Sweeney, “Generalizing data to provide anonymity when disclosing information,” in *PODS*, vol. 98, 1998, p. 188.

[28] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” technical report, SRI International, Tech. Rep., 1998.

[29] Y. Xu, T. Ma, M. Tang, and W. Tian, “A survey of privacy preserving data publishing using generalization and suppression,” *Appl. Math*, vol. 8, no. 3, pp. 1103–1116, 2014.

[30] L. Sweeney, “Guaranteeing anonymity when sharing medical data, the Datafly System.” in *Proceedings of the AMIA Annual Fall Symposium.* American Medical Informatics Association, 1997, p. 51.

[31] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient full-domain k-anonymity,” in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 49–60.

[32] B. Fung, K. Wang, and P. S. Yu, “Top-down specialization for information and privacy preservation,” in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. IEEE, 2005, pp. 205–216.

[33] B. Fung, K. Wang, L. Wang, and M. Debbabi, “A framework for privacy-preserving cluster analysis,” in *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on*. IEEE, 2008, pp. 46–51.

[34] B. Fung, K. Wang, and P. S. Yu, “Anonymizing classification data for privacy preservation,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 5, pp. 711–725, 2007.

[35] C. C. Aggarwal, “On k-anonymity and the curse of dimensionality,” in *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 2005, pp. 901–909.

[36] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 571–588, 2002.

[37] S. Zhong, Z. Yang, and T. Chen, “k-anonymous data collection,” *Information sciences*, vol. 179, no. 17, pp. 2948–2963, 2009.

[38] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira, “Efficient multidimensional suppression for k-anonymity,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 3, pp. 334–347, 2010.

[39] N. Matatov, L. Rokach, and O. Maimon, “Privacy-preserving data mining: A feature set partitioning approach,” *Information Sciences*, vol. 180, no. 14, pp. 2696–2720, 2010.

[40] M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[41] U. M. Fayyad and K. B. Irani, “Multi-interval discretization of continuous-valued attributes for classification learning,” in *IJCAI*, 1993, pp. 1022–1029.

[42] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, “Feedback on a publicly distributed database: the messidor database,” *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014.

[43] C. Kannas, K. Achilleos, Z. Antoniou, C. Nicolaou, C. Pattichis, I. Kalvari, I. Kirmitzoglou, and V. Promponas, “A workflow system for virtual screening in cancer chemoprevention,” in *IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*. IEEE, 2012, pp. 439–446.

[44] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[45] D. W. Aha, D. Kibler, and M. K. Albert, “Instance-based learning algorithms,” *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[46] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems?” *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[47] K. Hornik, C. Buchta, and A. Zeileis, “Open-source machine learning: R meets Weka,” *Computational Statistics*, vol. 24, no. 2, pp. 225–232, 2009.

[48] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[49] A. Antoniades, J. Keane, A. Aristodimou, C. Philipou, A. Constantinou, F. Tozzi, K. Kyriacou, A. Hadjisavvas, M. Loizidou, C. Demetriou, and C. Pattichis, “The effects of applying cell-suppression and perturbation to aggregated genetic data,” in *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 644–649.

[50] S. Garcia, J. Luengo, J. A. Sáez, V. Lopez, and F. Herrera, “A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 734–750, 2013.

[51] R. Ali, M. H. Siddiqi, and S. Lee, “Rough set-based approaches for discretization: a compact review,” *Artificial Intelligence Review*, vol. 44, no. 2, pp. 235–263, 2015.

[52] M. Richeldi and M. Rossotto, “Class-driven statistical discretization of continuous attributes,” in *European Conference on Machine Learning*. Springer, 1995, pp. 335–338.

[53] J. Dougherty, R. Kohavi, and M. Sahami, “Supervised and unsupervised discretization of continuous features,” in *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 194–202.

[54] S. Kotsiantis and D. Kanellopoulos, “Discretization techniques: A recent survey,” *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 47–58, 2006.

[55] S. Ramírez-Gallego, S. García, H. Mouriño-Talín, D. Martínez-Rego, V. Bolón-Canedo, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, “Data discretization: taxonomy and big data challenge,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 1, pp. 5–21, 2016.

[56] D. A. Zighed, S. Rabaséda, and R. Rakotomalala, “FUSINTER: a method for discretization of continuous attributes,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 03, pp. 307–326, 1998.

[57] X. Liu and H. Wang, “A discretization algorithm based on a heterogeneity criterion,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1166–1173, 2005.

[58] L.-Y. Wen, F. Min, and S.-Y. Wang, “A two-stage discretization algorithm based on information entropy,” *Applied Intelligence*, pp. 1–17, 2017.

[59] R. Kerber, “Chimerge: Discretization of numeric attributes,” in *Proceedings of the tenth national conference on Artificial intelligence*. Aaai Press, 1992, pp. 123–128.

[60] H. Liu and R. Setiono, “Feature selection via discretization,” *IEEE Transactions on knowledge and Data Engineering*, vol. 9, no. 4, pp. 642–645, 1997.

[61] F. E. Tay and L. Shen, “A modified chi2 algorithm for discretization,” *IEEE Transactions on knowledge and data engineering*, vol. 14, no. 3, pp. 666–670, 2002.

[62] L. Gonzalez-Abril, F. J. Cuberos, F. Velasco, and J. A. Ortega, “Ameva: An autonomous discretization algorithm,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5327–5332, 2009.

[63] L. A. Kurgan and K. J. Cios, “CAIM discretization algorithm,” *IEEE transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 145–153, 2004.

[64] A. Cano, D. T. Nguyen, S. Ventura, and K. J. Cios, “ur-CAIM: improved CAIM discretization for unbalanced and balanced data,” *Soft Computing*, vol. 20, no. 1, pp. 173–188, 2016.

[65] A. Cano, J. M. Luna, E. L. Gibaja, and S. Ventura, “Laim discretization for multi-label data,” *Information Sciences*, vol. 330, pp. 370–384, 2016.

[66] D. W. Scott, “Multivariate density estimation and visualization,” in *Handbook of computational statistics*. Springer, 2012, pp. 549–569.

[67] G. Schmidberger and E. Frank, “Unsupervised discretization using tree-based density estimation,” in *PKDD*, vol. 5. Springer, 2005, pp. 240–251.

[68] H. A. Sturges, “The choice of a class interval,” *Journal of the american statistical association*, vol. 21, no. 153, pp. 65–66, 1926.

[69] D. W. Walker and J. J. Dongarra, “Mpi: a standard message passing interface,” *Supercomputer*, vol. 12, pp. 56–68, 1996.

[70] L. Dagum and R. Menon, “Openmp: An industry-standard api for shared-memory programming,” *Computing in Science & Engineering*, no. 1, pp. 46–55, 1998.

[71] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017, r package version 1.6-8.

[72] H. Kim, *discretization: Data preprocessing, discretization for classification.*, 2012, r package version 1.0-1.

[73] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[74] B. A. McKinney *et al.*, “Machine learning for detecting gene–gene interactions: a review,” *Appl. bioinformatics*, vol. 5, no. 2, pp. 77–88, 2006.

[75] H. J. Cordell, “Detecting gene–gene interactions that underlie human diseases,” *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, June 2009.

[76] R. Upstill-Goddard *et al.*, “Machine learning approaches for the discovery of gene-–gene interactions in disease data,” *Briefings in Bioinformatics*, May 2012.

[77] M. D. Ritchie *et al.*, “Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer,” *The Amer. J. of Human Genetics*, vol. 69, no. 1, pp. 138–147, July 2001.

[78] Y. Chung *et al.*, “Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions,” *Bioinformatics*, vol. 23, no. 1, pp. 71–76, Jan. 2007.

[79] X.-Y. Lou *et al.*, “A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence,” *Amer. J. of Human Genetics*, vol. 80, no. 6, pp. 1125–1137, June 2007.

[80] J. Gui *et al.*, “A robust multifactor dimensionality reduction method for detecting Gene–Gene interactions with application to the genetic analysis of bladder cancer susceptibility,” *Ann. of Human Genetics*, vol. 75, no. 1, p. 20–28, 2011.

[81] D. Curtis, B. V. North, and P. C. Sham, “Use of an artificial neural network to detect association between a disease and multiple marker genotypes,” *Ann. of Human Genetics*, vol. 65, no. 1, p. 95–107, 2001.

[82] A. A. Motsinger *et al.*, “GPNN: power studies and applications of a neural network method for detecting gene–gene interactions in studies of human disease,” *BMC bioinformatics*, vol. 7, p. 39, 2006.

[83] A. A. Motsinger-Reif *et al.*, “Comparison of approaches for machine-learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology,” *Genetic epidemiology*, vol. 32, no. 4, pp. 325–340, May 2008.

[84] M. Waddell, D. Page, and J. Shaughnessy, Jr., “Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma,” in *Proc. of the 5th Int. Workshop on Bioinformatics*, ser. BIOKDD ’05. New York, NY, USA: ACM, 2005, pp. 21–28.

[85] S.-H. Chen *et al.*, “A support vector machine approach for detecting gene–gene interaction,” *Genetic Epidemiology*, vol. 32, no. 2, p. 152–167, 2008.

[86] A. Bureau *et al.*, “Identifying snps predictive of phenotype using random forests,” *Genetic Epidemiology*, vol. 28, no. 2, pp. 171–182, 2005.

[87] K. Lunetta *et al.*, “Screening large-scale association study data: exploiting interactions using random forests,” *BMC genetics*, vol. 5, no. 1, p. 32, 2004.

[88] M. Yoshida and A. Koike, “SNPInterForest: a new method for detecting epistatic interactions,” *BMC Bioinformatics*, vol. 12, no. 1, p. 469, Dec. 2011.

[89] D. F. Schwarz, I. R. König, and A. Ziegler, “On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data,” *Bioinformatics*, vol. 26, no. 14, pp. 1752–1758, July 2010.

[90] T. Kohonen, “The self-organizing map,” *Proc. of the IEEE*, vol. 78, no. 9, pp. 1464 –1480, Sept. 1990.

[91] P. Tamayo *et al.*, “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation,” *Proc. of the Nat. Academy of Sci.*, vol. 96, no. 6, pp. 2907–2912, Mar. 1999.

[92] C. Martin *et al.*, “Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification,” *Bioinformatics*, vol. 24, no. 14, pp. 1568–1574, July 2008.

[93] A. M. Newman and J. B. Cooper, “AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number,” *BMC Bioinformatics*, vol. 11, no. 1, p. 117, Mar. 2010.

[94] G. Skreti, E. Bei, and M. Zervakis, “Shape-influenced clustering of dynamic patterns of gene profiles,” in *Int. Conf. of the IEEE Eng. in Med. and Biol. Soc.*, San Diego, California, Sept. 2012.

[95] N. Chen and N. Marques, “An extension of self-organizing maps to categorical data,” in *Progress in Artificial Intell.*, ser. Lecture Notes in Computer Science, C. Bento, A. Cardoso, and G. Dias, Eds. Springer Berlin / Heidelberg, 2005, vol. 3808, pp. 304–313.

[96] M. Bahlo *et al.*, “Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20,” *Nature genetics*, vol. 41, no. 7, pp. 824–828, July 2009.

[97] J. R. Oksenberg *et al.*, “The genetics of multiple sclerosis: SNPs to pathways to pathogenesis,” *Nature Reviews Genetics*, vol. 9, no. 7, pp. 516–526, June 2008.

[98] G. René *et al.*, “Association of the HLA region with multiple sclerosis as confirmed by a genome screen using >10,000 SNPs on dna chips,” *J. of Molecular Medicine*, vol. 83, no. 6, pp. 486–494, 2005.

[99] J. Link *et al.*, “Two HLA class I genes independently associated with multiple sclerosis,” *J. of Neuroimmunology*, vol. 226, no. 1–2, pp. 172–176, Sept. 2010.

[100] D. Hafler *et al.*, “Risk alleles for multiple sclerosis identified by a genomewide study,” *The New England J. of medicine*, vol. 357, no. 9, pp. 851–862, Aug. 2007.

[101] A. Antoniades, “Discovering disease associated gene–gene interactions: A two snp interaction analysis framework,” *Ph.D. dissertation, University of Cyprus, Cyprus*, 2011.

[102] A. Antoniades *et al.*, “A computationally fast measure of epistasis for 2 SNPs and a categorical phenotype,” *IEEE EMBC*, vol. 2010, pp. 6194–6197, 2010.

[103] D. Brassat *et al.*, “Multifactor dimensionality reduction reveals gene–gene interactions associated with multiple sclerosis susceptibility in African Americans,” *Genes and Immunity*, vol. 7, no. 4, pp. 310–315, 2006.

[104] A. Motsinger *et al.*, “Complex gene–gene interactions in multiple sclerosis: a multifactorial approach reveals associations with inflammatory genes,” *Neurogenetics*, vol. 8, no. 1, pp. 11–20, 2007.

[105] J. H. Moore and S. M. Williams, “Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis,” *BioEssays*, vol. 27, no. 6, p. 637–646, 2005.

[106] M. Emily, “A survey of statistical methods for gene-gene interaction in case-control genome-wide association studies,” *Journal de la Societe Française de Statistique*, vol. 159, no. 1, pp. 27–67, 2018.

[107] T. Hu, N. A. Sinnott-Armstrong, J. W. Kiralis, A. S. Andrew, M. R. Karagas, and J. H. Moore, “Characterizing genetic interactions in human disease association studies using statistical epistasis networks,” *BMC bioinformatics*, vol. 12, no. 1, p. 364, 2011.

[108] T. Hu, Y. Chen, J. W. Kiralis, R. L. Collins, C. Wejse, G. Sirugo, S. M. Williams, and J. H. Moore, “An information-gain approach to detecting three-way epistatic interactions in genetic association studies,” *Journal of the American Medical Informatics Association*, 2013.

[109] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, “A survey about methods dedicated to epistasis detection,” *Frontiers in Genetics*, vol. 6, p. 285, 2015.

[110] Y. M. Cho, M. D. Ritchie, J. H. Moore, J. Y. Park, K. U. Lee, H. D. Shin, H. K. Lee, and K. S. Park, “Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus,” *Diabetologia*, vol. 47, no. 3, p. 549–554, 2004.

[111] D. Brassat, A. A. Motsinger, S. J. Caillier, H. A. Erlich, K. Walker, L. L. Steiner, B. a. C. Cree, L. F. Barcellos, M. A. Pericak-Vance, S. Schmidt, S. Gregory, S. L. Hauser, J. L. Haines, J. R. Oksenberg, and M. D. Ritchie, “Multifactor dimensionality reduction reveals gene–gene interactions associated with multiple sclerosis susceptibility in african americans,” *Genes and Immunity*, vol. 7, no. 4, pp. 310–315, 2006.

[112] A. S. Andrew, M. R. Karagas, H. H. Nelson, S. Guarrera, S. Polidoro, S. Gamberini, C. Sacerdote, J. H. Moore, K. T. Kelsey, and E. Demidenko, “DNA repair polymorphisms modify bladder cancer risk: a multi-factor analytic strategy,” *Human heredity*, vol. 65, no. 2, p. 105–118, 2008.

[113] O.-Y. Fu, H.-W. Chang, Y.-D. Lin, L.-Y. Chuang, M.-F. Hou, and C.-H. Yang, "Breast cancer-associated high-order snp-snp interaction of cxcl12/cxcr4-related genes by an improved multifactor dimensionality reduction (mdr-er)," *Oncology reports*, vol. 36, no. 3, pp. 1739–1747, 2016.

[114] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Advances in knowledge discovery and data mining," U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, ch. Fast Discovery of Association Rules, pp. 307–328.

[115] D. Tian, A. Gledson, A. Antoniades, A. Aristodimou, N. Dimitrios, R. Sahay, J. Pan, S. Stivaros, G. Nenadic, X.-j. Zeng, *et al.*, "A bayesian association rule mining algorithm," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 3258–3264.

[116] S. Uppu, A. Krishna, and R. P. Gopalan, "Rule-based analysis for detecting epistasis using associative classification mining," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 4, no. 1, p. 12, 2015.

[117] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[118] H. Chipman and R. Tibshirani, "Hybrid hierarchical clustering with applications to microarray data," *Biostatistics*, vol. 7, no. 2, pp. 286–301, Apr. 2006.

[119] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, Mar. 2012.

[120] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Technical Report 1996-77, February 1996.

[121] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[122] K. Kira, L. A. Rendell, *et al.*, "The feature selection problem: Traditional methods and a new algorithm," in *Aaai*, vol. 2, 1992, pp. 129–134.

[123] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[124] A. Chaturvedi, P. E. Green, and J. D. Caroll, "K-modes clustering," *Journal of classification*, vol. 18, no. 1, pp. 35–55, 2001.

[125] E. Dardiotis, E. Panayiotou, V. Siokas, A.-M. Aloizou, K. Christodoulou, A. Hadjisavvas, M. Pantzaris, N. Grigoriadis, G. M. Hadjigeorgiou, and T. Kyriakides, "Gene variants of adhesion molecules predispose to ms: A case-control study," *Neurology Genetics*, vol. 5, no. 1, p. e304, 2019.

[126] A. Antoniades *et al.*, "A computationally fast measure of epistasis for 2 snps and a categorical phenotype," in *Eng. in Medicine and Biology Soc. (EMBC), 2010 Annu. Int. Conf. of the IEEE*, 2010, pp. 6194–6197.

[127] M. Kutmon, M. P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A. R. Pico, and C. T. Evelo, "Pathvisio 3: An extendable pathway analysis toolbox," *PLOS Computational Biology*, vol. 11, no. 2, pp. 1–13, 02 2015.

[128] D. Zamar, B. Tripp, G. Ellis, and D. Daley, "Path: a tool to facilitate pathway-based genetic association analysis," *Bioinformatics*, vol. 25, no. 18, pp. 2444–2446, 07 2009.

[129] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[130] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[131] Z. Zhang, M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda, H. Goyal, *et al.*, "Opening the black box of neural networks: methods for interpreting neural network models in clinical applications," *Annals of translational medicine*, vol. 6, no. 11, 2018.

Appendix A

Publications

A.1 Journals

A.1.1 Published

1. **A. Aristodimou**, A. Antoniades, C. Georgousopoulos, N. Forgó, A. Gledson, P. Hasapis, C. Vandeleur, K. Perakis, R. Sahay, M. Mehdi, C. A. Demetriou, M.-P. F. Strippoli, V. Giotaki, M. Ioannidi, D. Tian, F. Tozzi, J. Keane, and C. Pattichis, "Advancing clinical research by semantically interconnecting aggregated medical data information in a secure context," Health and Technology, vol. 7, no. 2, pp. 223-240, Nov 2017

2. **A. Aristodimou**, A. Antoniades, and C. S. Pattichis, "Privacy preserving data publishing of categorical data through k-anonymity and feature selection," Healthcare Technology Letters, vol. 3, no. 1, pp. 16-21, 2016.

A.1.2 Submitted

- **A. Aristodimou**, A. Diavastos and C. Pattichis, "A Supervised Density Based Discretization Algorithm for Big Data Classification Tasks in the Medical Domain", Computer Methods and Programs in Biomedicine (CMPB), July 2019

- **A. Aristodimou** *et al*, "A Framework for Efficient n-Way Interaction Testing in Case/Control Studies", BMC Bioinformatics, to be submitted in December 2019

A.2 Conference Papers

A.2.1 Published

3. N. Marios, K. Neokleous, **A. Aristodimou**, I. Constantinou, Z. Antoniou, E. C. Schiza, C. S. Pattichis, and C. N. Schizas. "Electronic health record application support service enablers." In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 1401-1404.

4. D. Tian, A. Gledson, A. Antoniades, **A. Aristodimou**, N. Dimitrios, R. Sahay, J. Pan, S. Stivaros, G. Nenadic, XJ. Zeng, J. Keane, "A Bayesian Association Rule Mining Algorithm," in 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 3258-3264.

5. S. Ratnesh, D. Ntalaperas, E. Kamateri, P. Hasapis, O. D. Beyan, MP. F. Strippoli, C. A. Demetriou , T. Stavropoulou, M. Brochhausen, K. Tarabanis, T. Bouras, D. Tian, **A. Aristodimou**, A. Antoniades, C. Georgousopoulos, M. Hauswirth, S. Decker. ”An ontology for clinical trial data integration.” In 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 3244-3250.

6. P. Hasapis, D. Ntalaperas, C. Kannas, **A. Aristodimou**, D. Alexandrou, T. Bouras, C. Georgousopoulos, A. Antoniades, C. Pattichis, A. Constantinou, ”Molecular clustering via knowledge mining from biomedical scientific corpora,” in 13th IEEE International Conference on BioInformatics and BioEngineering, 2013, pp. 1-5.

7. **A. Aristodimou**, A. Antoniades, and C. Pattichis, ”Clustering subjects in genetic studies with Self Organizing Maps,” in 2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE), 2012, pp. 546-551.

8. A. Antoniades, C. Georgousopoulos, N. Forgo, **A. Aristodimou**, F. Tozzi, P. Hasapis, K. Perakis, T. Bouras, D. Alexandrou, E. Kamateri, E. Panopoulou, C. Pattichis. ”Linked2Safety: A secure linked data medical information space for semantically-interconnecting EHRs advancing patients’ safety in medical research,” in 2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE), 2012, pp. 517-522.

9. A. Antoniades, J. Keane, **A. Aristodimou**, C. Philipou, A. Constantinou, C. Georgousopoulos, F. Tozzi, K. Kyriacou, A. Hadjisavvas, M. Loizidou, C. Demetriou, C. Pattichis ”The effects of applying cell-suppression and perturbation to aggregated genetic data,” in 2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE), 2012, pp. 644-649.

10. A. Athos, L. Loizou, **A. Aristodimou**, and C. S. Pattichis. ”A binary format for genetic data designed for large whole genome studies that enable both marker and strand based analyses.” In 2008 8th IEEE International Conference on BioInformatics and BioEngineering, 2008, pp. 1-4.